

Understanding the germline and cancer variome through pan-cancer analysis using HIVE

Raja Mazumder mazumder@gwu.edu

Assoc. Prof. Biochemistry and Molecular Medicine

public-HIVE Project Lead

Director, The McCormick Genomic & Proteomic Center

GWU



High-performance Integrated Virtual Environment (HIVE)

HIVE - Google Chrome
https://hive.biochemistry.gwu.edu/dna.cgi?cmd=main

Main | Contact Help Login Register News

Main
About HIVE
Tutorial
Publications
Tools
People
Employment

Welcome to HIVE
High-performance Integrated Virtual Environment

NGS

HIVE in conjunction with the Genomic Work Group at the FDA is planning a public workshop
"Next-generation sequencing (NGS) technology, data formats standardization and promotion of interoperability protocols"

HIVE is a cloud-based environment optimized for the storage and analysis of extra-large data, like Next Generation Sequencing data, Mass Spectroscopy files, Confocal Microscopy Images and others.

HIVE uses a variety of advanced scientific and computational visualization graphics, to get the **MOST** from your HIVE experience you must use a supported browser. These include Internet Explore 8.0 or higher (Internet Explorer 9.0 is recommended), Google Chrome, Mozilla Firefox and Safari.

A few exemplary analytical outputs are displayed below for your enjoyment. But before you can take advantage of all that HIVE has to offer and create these objects for yourself, you'll need to register.

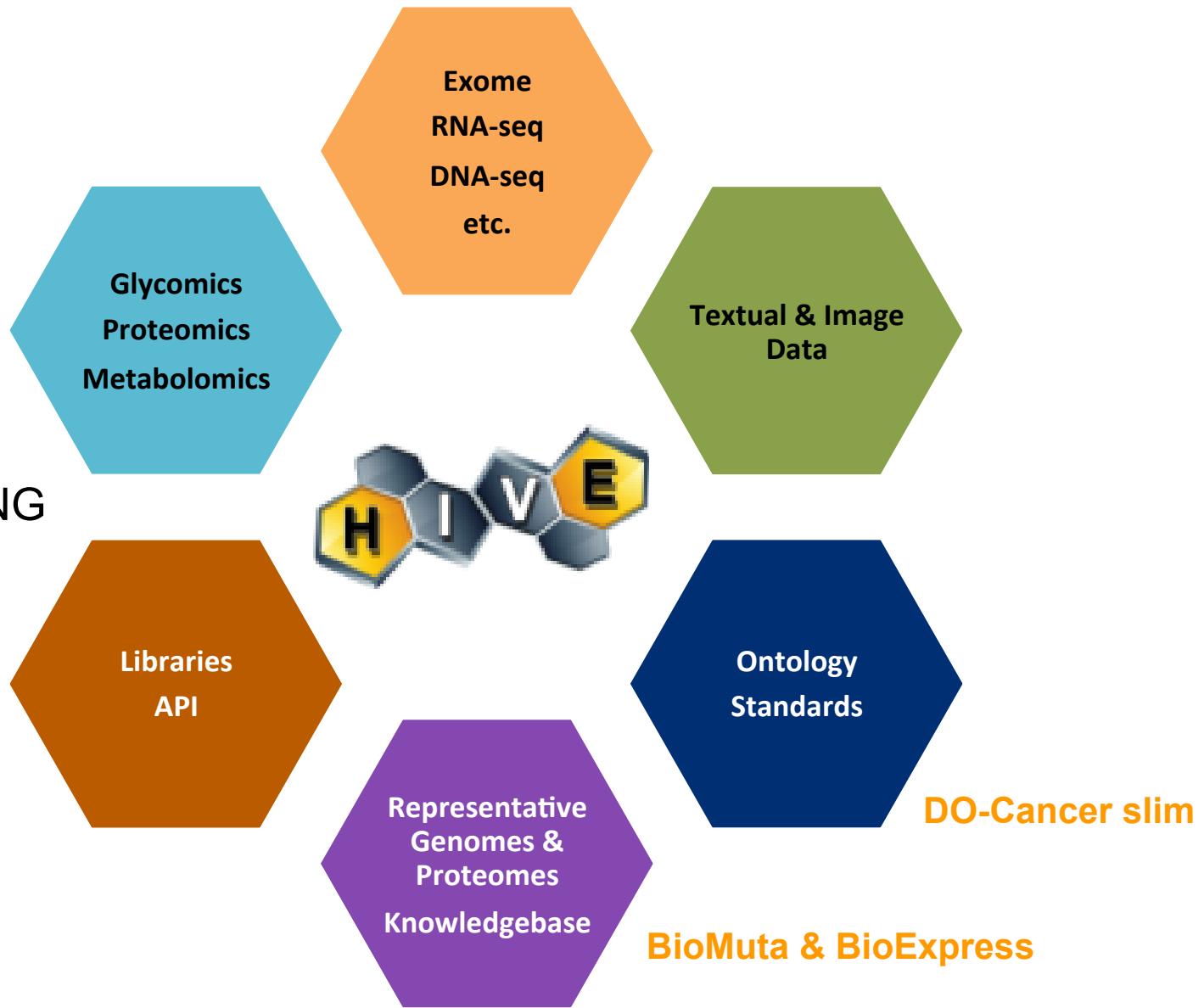
September 24 and 25, 2014, from 8:30 a.m. to 4:30 p.m.

2 portals: Public portal (GWU) & a FDA only portal

mazumder@gwu.edu | vahan.simonyan@fda.hhs.gov

HIVE + Community tools & databases

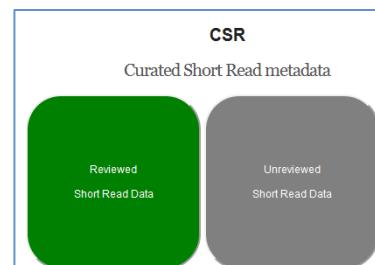
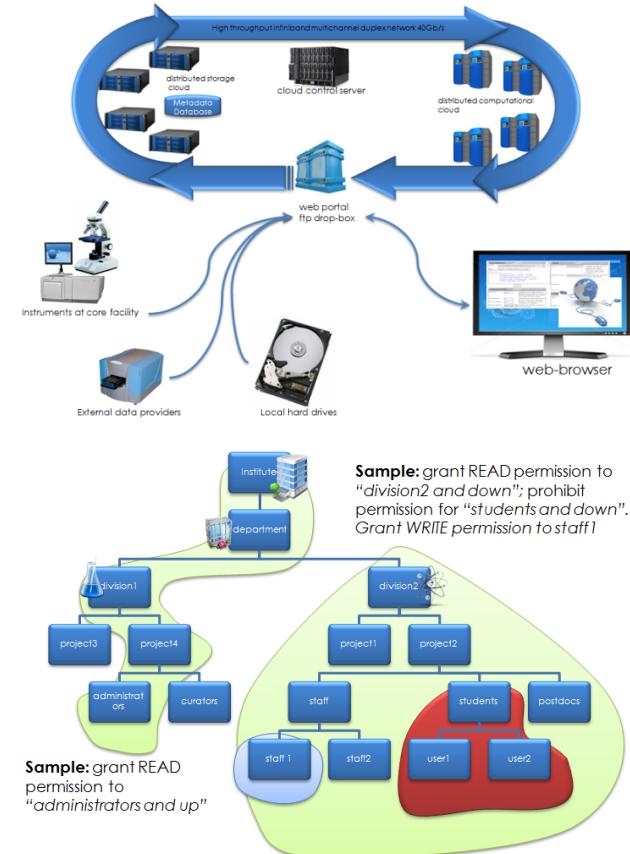
INTEGRATED
BIG DATA
STORAGE, SHARING
AND ANALYSIS
PLATFORM



References + Standards in collaboration with community

HIVE key features and focus

-  installations: In-house HPC, Amazon, HIVE-in-a-box
- Compute speed
- Automatic distributed storage and computing
- Security features (FDA/ patient data)
- Granular sharing capabilities; Traceability
- Knowledgebases:
Biocuration of content



BioMutate v1.0

New version available - BioMutate v2.0(beta)

BioExpress v1.0

Loss and gain of N-linked glycosylation sites due to single-nucleotide variations

- **N-X-(S/T)** sequence motif required for N-glycosylation of proteins. N: asparagine; S: serine; T: threonine; X: cannot be proline.
- Not all **N-X-(S/T)** sequences are glycosylated.

LOSS	GAIN
AAU (N -X-T) ↓mutation AAG (K -X-T)	AAG (K -X-T) ↓mutation AAU (N -X-T)

Somatic mutations
Or polymorphisms

Proteome-Wide Analysis of Single-Nucleotide Variations in the N-Glycosylation Sequon of Human Genes 2012

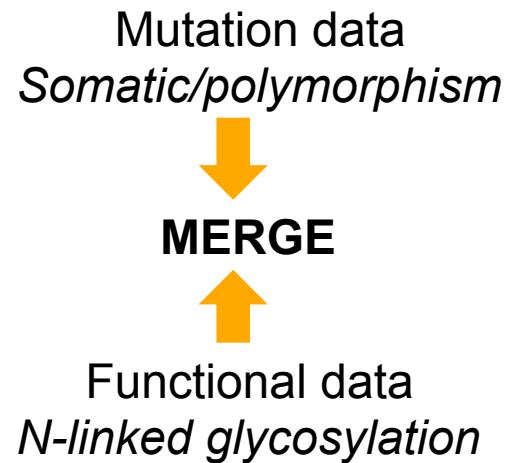
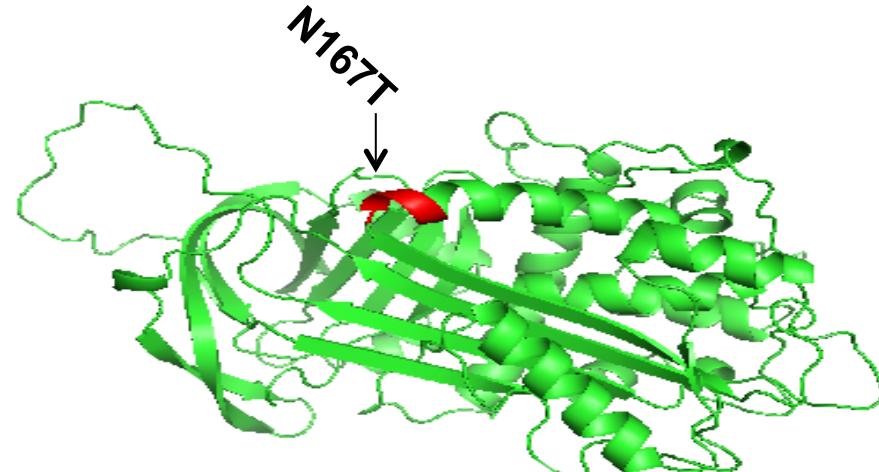
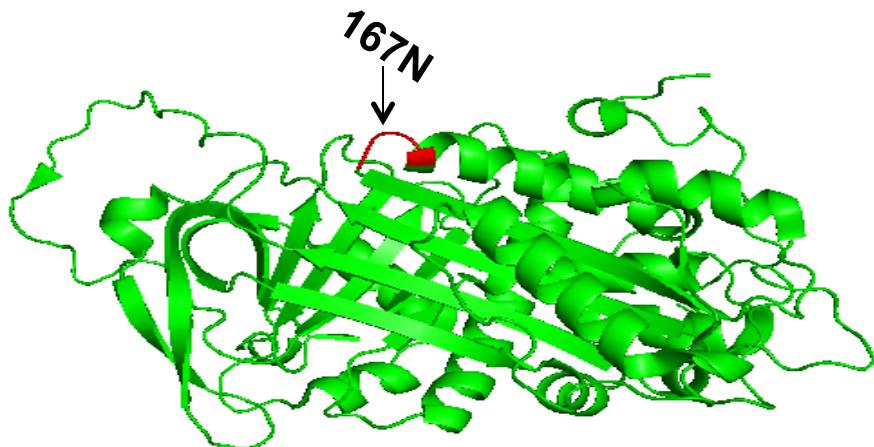
Raja Mazumder^{1,*}, Krishna Sudeep Morampudi^{2,3*}, Mona Motwani¹, Sona Vasudevan³, Radoslav Goldman^{2,3}

Structure-based Comparative Analysis and Prediction of N-linked Glycosylation Sites in Evolutionarily Distant Eukaryotes

Phuc Vinh Nguyen Lam^{1,3}, Radoslav Goldman², Konstantinos Karagiannis³, Tejas Narsule³, Vahan Simonyan⁴, Valerii Soika⁴, Raja Mazumder^{3,*}

SNVDis: A Proteome-wide Analysis Service for Evaluating nsSNVs in Protein Functional Sites and Pathways

Konstantinos Karagiannis¹, Vahan Simonyan², Raja Mazumder^{1,*}



2013	ID	Position	Function
	3KCG	167	N-gly

ID	Position	Mutation
P01008	167	N->T

N-linked glycosylation

- Significant number of proteins that gained or lost the glycosylation are involved in kinase activity, immune response, and blood coagulation.
- ~70% of all asparagines of NXS/T motif involved in N-glycosylation are localized in the loop/turn conformation in all species studied.
- Using machine learning methods based on rules we can predict with 93% accuracy if a particular site will be glycosylated.
- 108 human proteins with structures and 2247 proteins without structures that have gained glycosylation sites.

Vogt G, Vogt B, Chuzhanova N, Julenius K, Cooper DN, Casanova JL. Gain-of-glycosylation mutations. Curr Opin Genet Dev. 2007

Nicolaou N, Margadant C, et al. Gain of glycosylation in integrin α3 causes lung disease and nephrotic syndrome. J Clin Invest. 2012.

Radivojac P, Baenziger PH, et al. Gain and loss of phosphorylation sites in human cancer. Bioinformatics. 2008 Aug 15;24(16):i241-7.

Proteome-wide variation impact analysis

HIVE GW SCHOOL OF MEDICINE AND HEALTH SCIENCES High-performance Integrated Virtual Environment (HIVE) Welcome to HIVE We welcome you to HIVE: High-performance Integrated Virtual Envi... Login Register Guest

Home Tools Data Courses People Links Contact About

Select datasets

help nsSNV sources

<input checked="" type="checkbox"/>	Name	Version	Link
<input checked="" type="checkbox"/>	UniProtKB	date of access 04/14/2013	www.uniprot.org/manual/variant
<input checked="" type="checkbox"/>	dbSNP	build 137	www.ncbi.nlm.nih.gov/projects/SNP/
<input checked="" type="checkbox"/>	COSMIC	version 64	www.sanger.ac.uk/genetics/CGP/cosmic/
<input checked="" type="checkbox"/>	NCI-60	-	www.sanger.ac.uk/genetics/CGP/NCI60/
<input checked="" type="checkbox"/>	TCGA-Prostate-Control	-	hive.biochemistry.gwu.edu/usr/satti/SRARecords_Curated.PHP
<input checked="" type="checkbox"/>	TCGA-Prostate-Cancer-	-	hive.biochemistry.gwu.edu/usr/satti/SRARecords_Curated.PHP
<input checked="" type="checkbox"/>	nextProt	release 2012-02-12	www.nextprot.org

help Proteome Sources Upload proteome

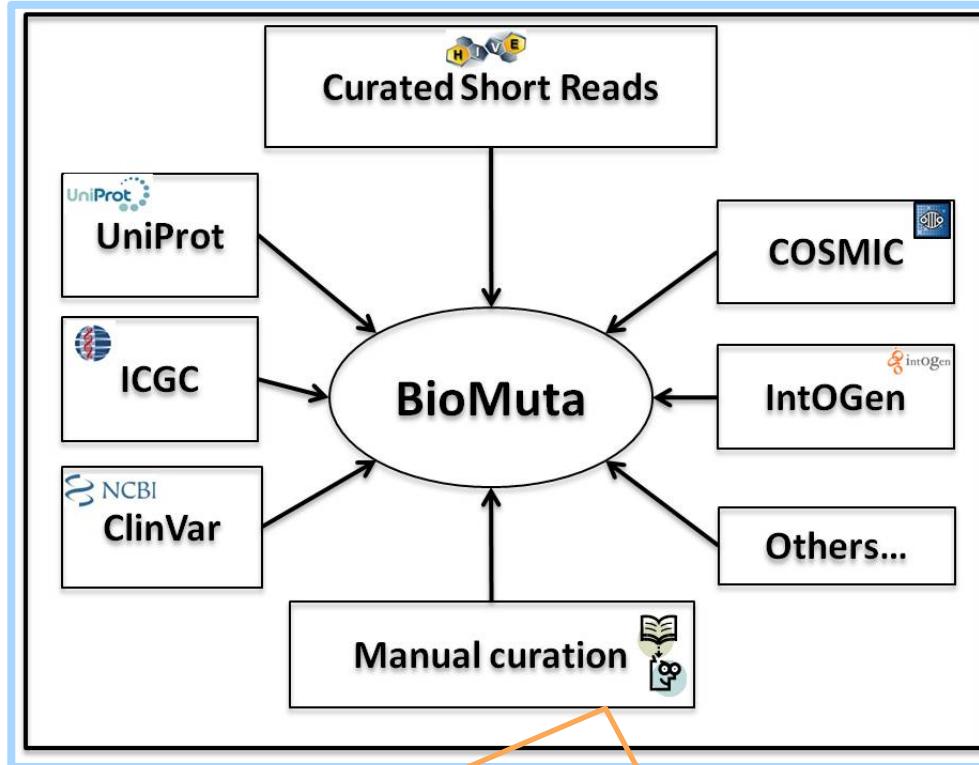
<input checked="" type="checkbox"/>	Source	Version
<input checked="" type="checkbox"/>	UniProtKB/Swiss-Prot	date of access 04/14/2013

Proteome-wide analysis allows identification of experimental targets

BioMuta

- BioMuta integrates all cancer related non-synonymous mutations.
- More than 1.4 million evidence tagged nsSNVs have been integrated into database (largest curated dataset that we know of)
- Disease Ontology nomenclature applied by BioMuta provides more accurate disease description allowing for pan-cancer analysis
- 381 DO terms and 54 DO slim terms are included in BioMuta

Databases of Variants



Search terms	Total articles ^a	Positive articles ^b
SNP, biomarker, cancer	702	60
Biomarker, cancer, single-nucleotide-polymorphism	1986	43
Polymorphism, biomarker, cancer	5215	20
SNP, exon, cancer	394	16
Genes ^c , cancer,.snp	20	4
Total		143 ^d

Incomplete variation information

Gene/Protein accession/Gene name	Genomic coordinates	Variation Gene/Protein (position)	Cancer definition	PMID	source
NM_130800.2 O00255 MEN1	64575133-64575133 (chr 11)	C A (1193); G V (230)	Lung, upper right lung, mucous cell, carcinoma	---	COSMIC
---- P40637 TP53	chr17:7579866	---; Q239L	Sporadic cancer	14660012	UniProt
NM_77692.4 ---- TP53	Chr17(7757534)	---; ---	Cancer	1791428	Manual
NM_533167.1 O20147 ---	----	2133(T G); G703P	Pancreas	31229574	IntOGen

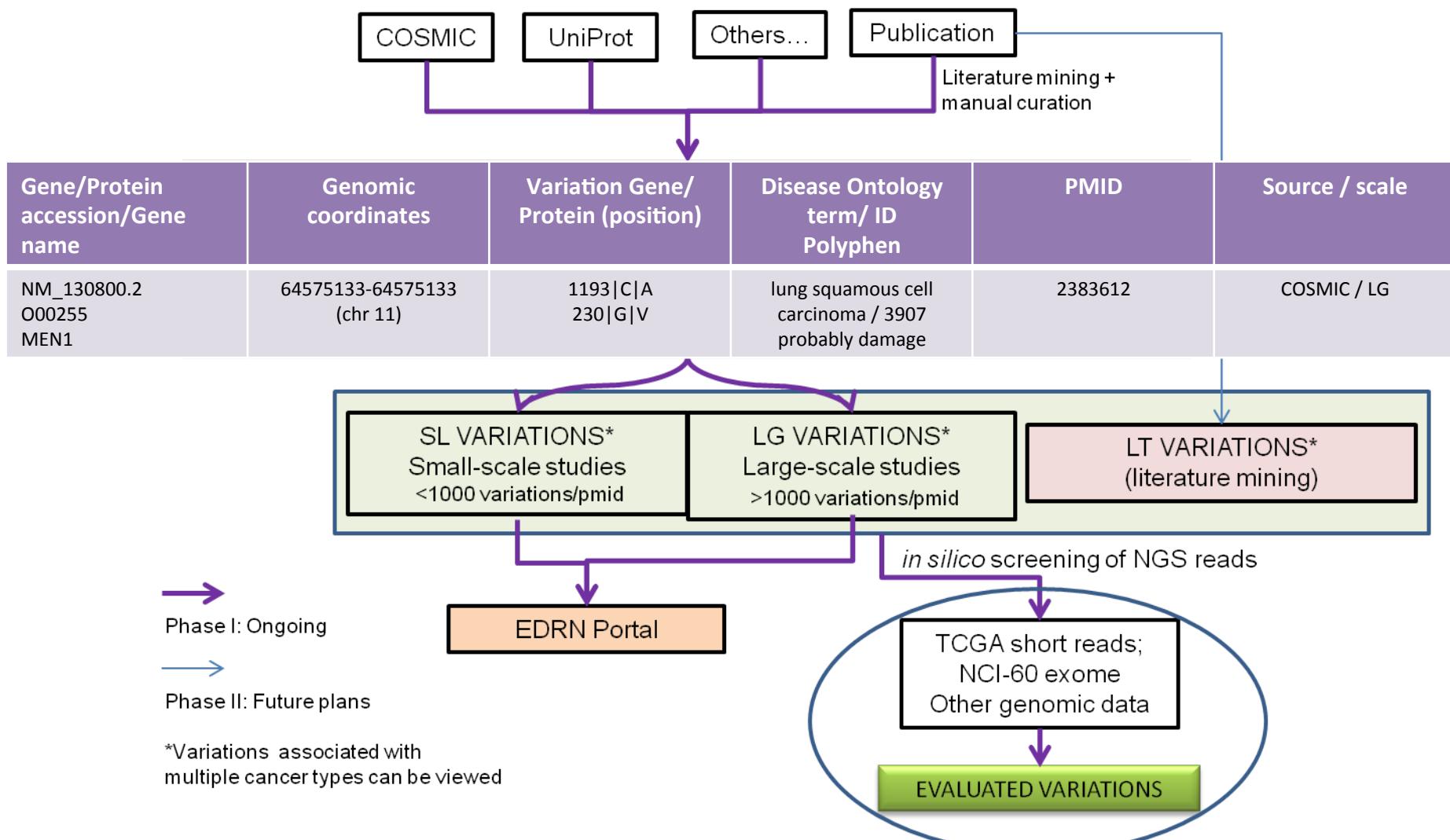
Disease Ontology (DO) and DO slim

Source	Original cancer term	DOID / DO term	DO_slim
IntOGen	Pancreas	DOID:1793 / pancreatic cancer	
TCGA	Pancreatic adenocarcinoma [PAAD]	DOID:4074 / pancreas adenocarcinoma	
COSMIC	pancreas,NS,carcinoma, acinar_carcinoma	DOID:5742 / pancreatic acinar cell adenocarcinoma	DOID:1793 / pancreatic cancer
UniProt	Pancreatic cancer	DOID:1793 / pancreatic cancer	

Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012 Jan;40(Database issue):D940-6.

- DO provides accurate disease description for all cancer terms
- DO slim group several terms of DO
- DO slim is easy for later analysis
- Common Data Elements (CDE) EDRN data portal

BioMuta workflow



BioMuta content

Scale	nsSNVs
LG	1,365,645
SM	44,378

Coverage	Counts
PMID	1562
DO term/ID	381
DO slim	54
UniProtKB	15,366

Source	ICGC	IntOGen	COSMIC	CSR-TSA	ClinVar	UniProtKB	Manual	Total
nsSNVs	519,780	432,729	418,488	31,979	4,590	2,281	176	1,410,023
PMIDs	1319	1	1	1	3	118	182	1562

LG: large scale
SM: small scale
DO: disease ontology
DO slim: disease ontology slim

Early
Detection
Research
Network

BioMuta SNV table

UCSC Genome Bioinformatics



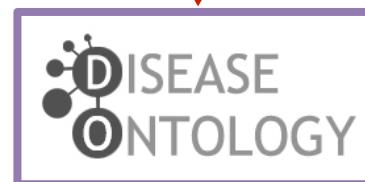
Source

DOWNLOAD

1450 Results for MUC16 in BioMuta v2.0 ▾																
UniProtKB	A	Gene	Accession	SNV position	Pos(N)	Ref(N)	Var(N)	Pos(A)	Ref(A)	Var(A)	Polyphen Pred.	PMID	Disease Ontology Name	Source	Status	Comments
Q8WXI7		MUC16	NM_02...	chr19:9091786-90...	233	G	T	10	S	*	-	20111...	DOID:1324 / lung cancer	IntOGen	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091771-90...	248	C	T	15	R	H	benign	20393...	DOID2871 / endometri...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091756-90...	263	C	T	20	G	E	benign	20393...	DOID3907 / lung squa...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091636-90...	383	G	A	60	P	L	benign	20393...	DOID3907 / lung squa...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091636-90...	383	G	A	60	P	L	benign	20111...	DOID:1324 / lung cancer	IntOGen	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091606-90...	413	G	A	70	S	L	benign	20393...	DOID2871 / endometri...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091579-90...	440	A	G	79	L	S	possibly dama...	24467...	DOID:3459 / breast car...	CSR-T...	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091578-90...	441	C	A	79	L	F	possibly dama...	20393...	DOID1996 / rectum ad...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091544-90...	475	A	G	91	S	P	benign	20393...	DOID3459 / breast car...	ICGC	LG	✉
Q8WXI7		MUC16	NM_02...	chr19:9091544-90...	475	A	G	91	S	P	benign	20111...	DOID:1612 / breast ca...	IntOGen	LG	✉

Swiss-Prot

RefSeq



EDRN data portal

NATIONAL CANCER INSTITUTE National Cancer Institute U.S. National Institutes of Health | www.cancer.gov

 Early Detection Research Network
Biomarkers: the key to early detection

DCP Division of Cancer Prevention [Log in](#)

[Home](#) [About EDRN](#) **Biomarkers** [Protocols](#) [Science Data](#) [Publications](#) [Resources](#) [Specimens](#)

You are here: [Home](#) / [Biomarkers](#) / CA125

CA125

[Search Site](#)

Basics [Organs](#) [Studies](#) [Publications](#) [Resources](#)

Aliases:
This biomarker is also known as:
Ovarian cancer-related tumor marker CA125, CA125 ovarian cancer antigen, MUC-16, FLJ14303, Ovarian carcinoma antigen CA125, CA-125, Mucin 16, mucin 16, cell surface associated, MUC16, Mucin-16,

ATTRIBUTES
QA State: Accepted
Type: Protein
Short Name:

News
The final report of the 2013 Cancer Biomarkers Bioinformatics Workshop is now available.

Announcement 06/26/2014

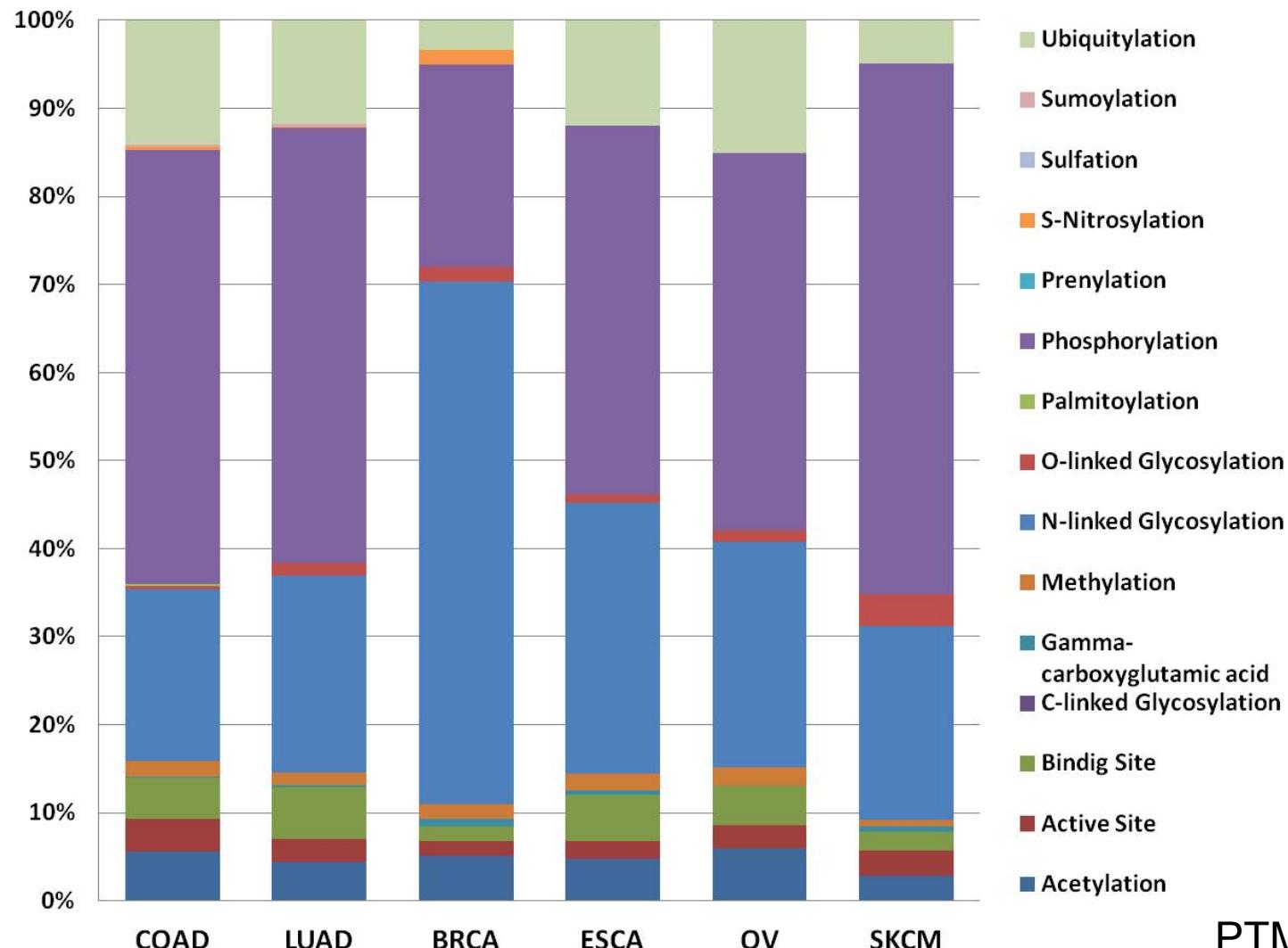
Please [click here to register](#) for the 9th EDRN Scientific Workshop from September 8-10, 2014, in Bethesda, Maryland. The meeting registration page also has agendas and hotel reservation information.

[View in BioMutu](#)

Description...
MUC16 (CA125) is a highly glycosylated sialomucin that is expressed on epithelial cell surface, especially on ovarian cancer cells. MUC16 is anchored to the epithelium by a transmembrane

[Network Consulting Team](#)
[Informatics](#)
[Collaborative Groups](#)
[Secure Site](#)
[Public, Patients, Advocates](#)
[Funding Opportunities](#)
[Sites](#)
[Member Directory](#)

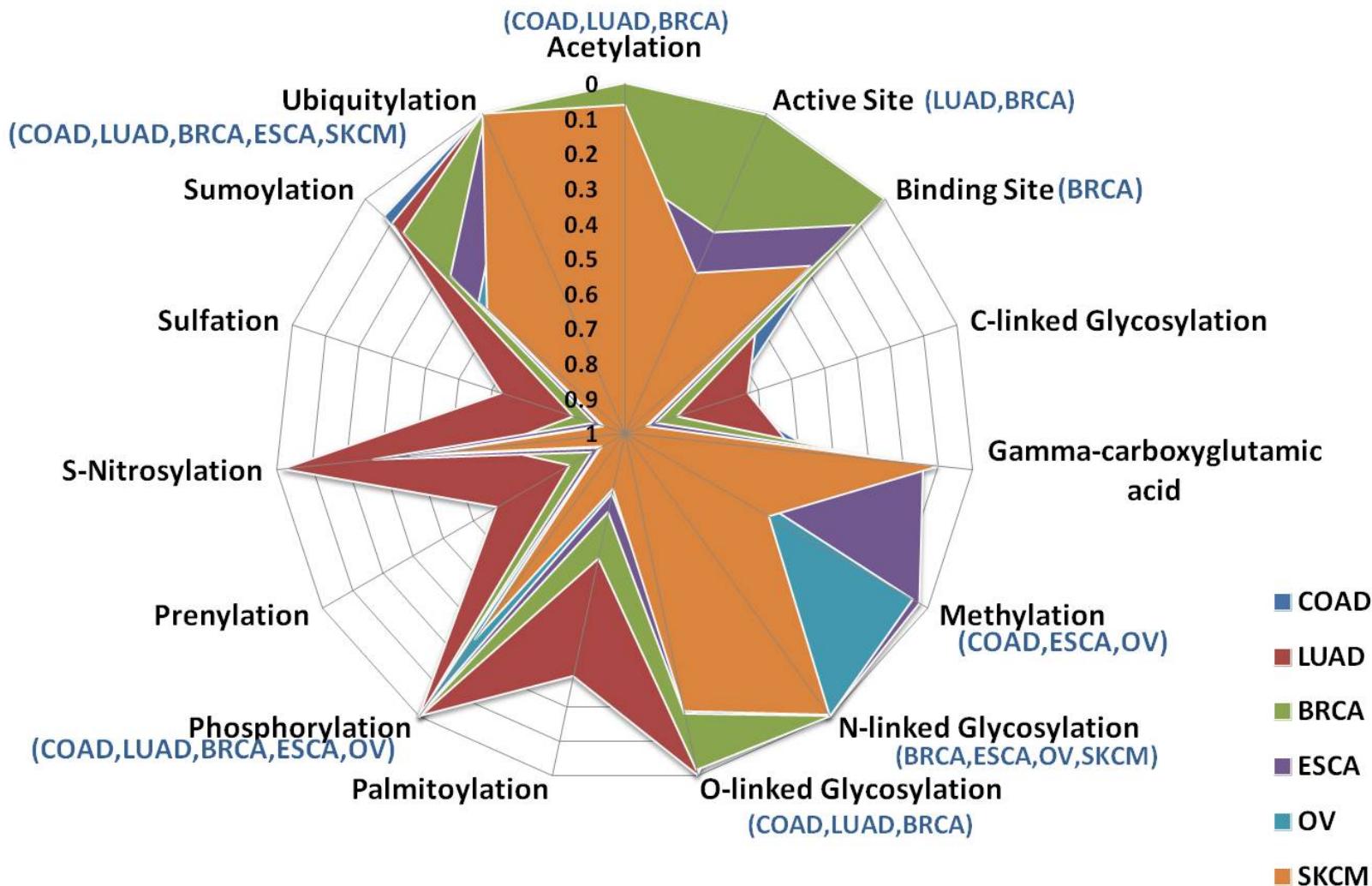
Effect of mutations on PTMs



Wu TJ, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, Mazumder R. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database (Oxford). 2014 Mar 25;2014:bau022.

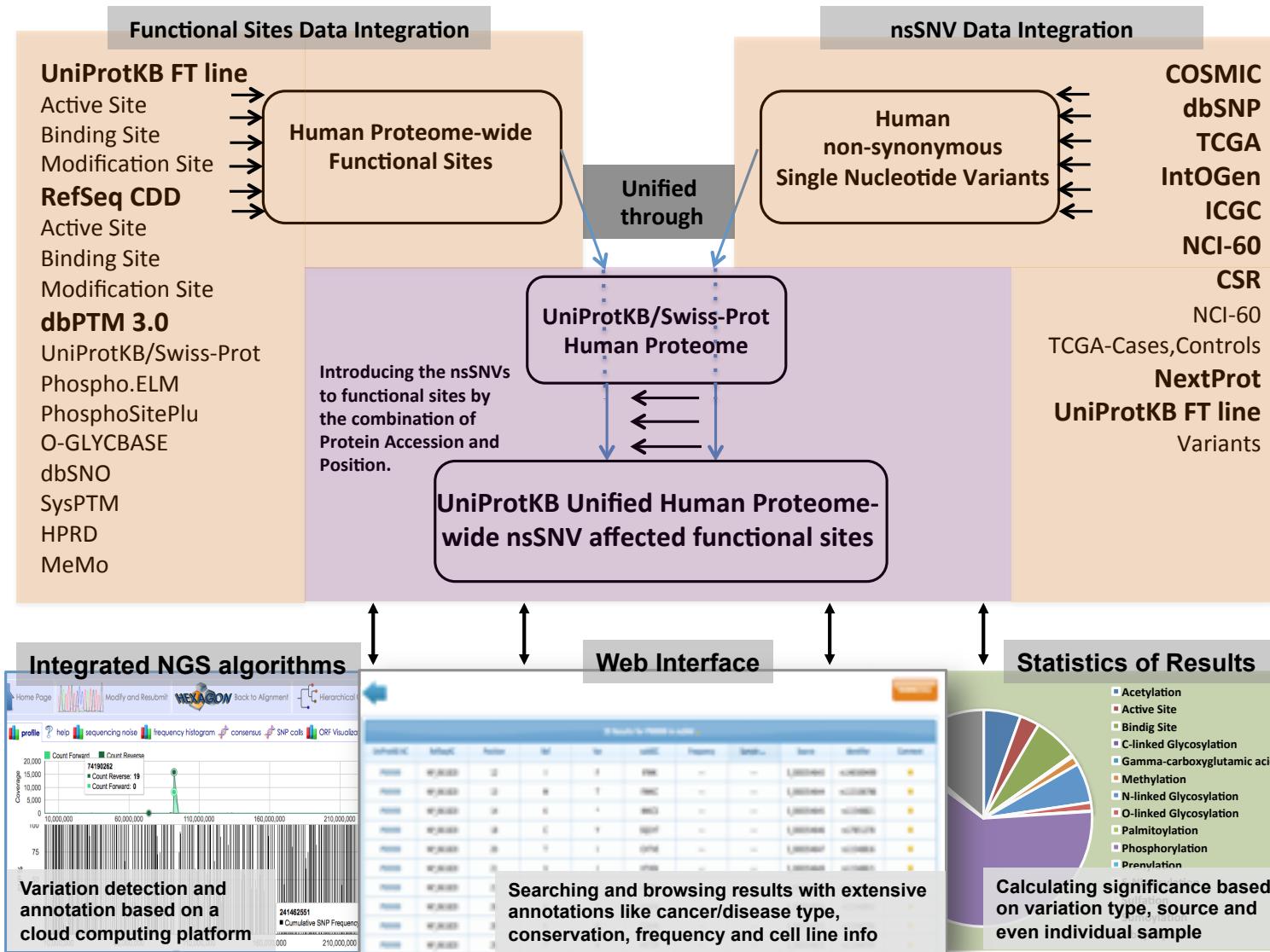
PTM: post-translational modification

PTMs significantly affected by cancers



PTM: post-translational modification

nSNV impact analysis workflow



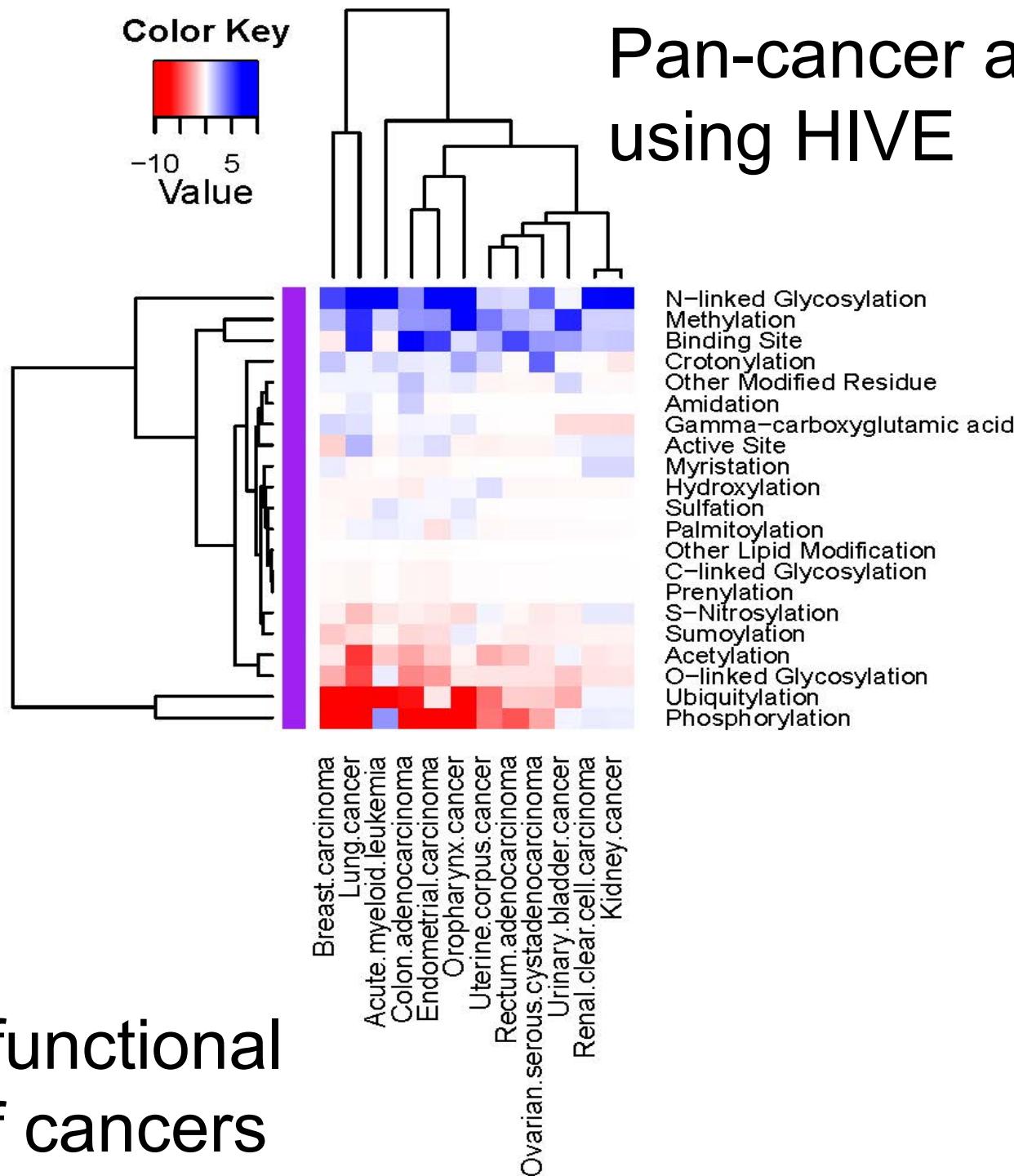
Summary of nsSNV affected functional sites obtained through proteome-wide survey

	COSMIC	TCGA	ICGC	IntOGen	CSR	NCI60	dbSNP	UniProt	nextProt	Total	Expected	Difference	p-value
Acetylation	185	322	198	134	15	15	351	65	190	863	1242.877056	-379.8770564	2.30E-30
Active Site	487	1051	506	379	27	40	811	247	541	2385	2214.980703	170.0192966	1.83E-04
Amidation	4	1	0	1	1	1	4	1	2	9	5.572089069	3.427910931	1.12E-01
Binding Site	3877	8380	4010	3054	253	451	6395	1549	4005	18681	17131.61374	1549.386262	4.81E-32
C-linked Glycosylation	1	1	2	1	0	0	0	0	0	4	10.84298413	-6.842984134	1.68E-02
Crotonylation	26	24	15	16	0	0	10	0	7	52	30.72178838	21.27821162	2.89E-04
Gamma-carboxyglutamic acid	5	13	6	7	0	0	11	20	22	36	12.34895415	23.65104585	3.46E-08
Hydroxylation	13	13	6	7	0	1	17	2	9	41	42.61895153	-1.618951527	4.42E-01
Methylation	78	109	63	71	8	3	61	32	59	224	102.4059613	121.5940387	2.24E-25
Myristylation	4	4	1	1	1	0	2	0	4	9	9.788805121	-0.788805121	4.85E-01
N-linked Glycosylation	732	1705	843	656	84	125	1997	207	859	4372	2435.15352	1936.84648	4.53E-273
O-linked Glycosylation	28	58	35	19	4	8	108	22	52	205	383.8717577	-178.8717577	8.62E-24
Other Lipid Modification	0	1	0	0	0	0	0	0	0	1	2.861343035	-1.861343035	2.21E-01
Other Modified Residue	14	32	16	10	2	1	24	7	16	70	46.23327957	23.76672043	6.75E-04
Palmitoylation	3	10	5	3	0	0	8	3	3	20	28.16163935	-8.161639348	6.88E-02
Phosphorylation	1630	3764	1871	1272	180	257	3917	436	1835	9383	11706.50734	-2323.507343	1.28E-110
Prenylation	0	1	0	0	0	0	2	0	0	3	10.99358114	-7.993581136	4.94E-03
S-Nitrosylation	15	19	13	11	0	0	43	7	20	75	113.7007364	-38.70073641	7.20E-05
Sulfation	1	4	5	1	0	0	4	5	5	13	12.80074516	0.199254842	5.15E-01
Sumoylation	10	21	7	8	0	2	13	0	8	48	95.02670818	-47.02670818	7.41E-08
Ubiquitylation	449	798	502	302	29	52	841	104	395	2055	3395.811795	-1340.811795	1.18E-136

Variant type based comparison of nsSNV's impact on different types of functional sites

	Total	+/- (Somatic) ¹	Not in dbSNP (Somatic)	+/- (Germline) ²	In dbSNP (Germline)
Acetylation	2.30E-30	-	4.39E-17	-	3.82E-15
Active Site	1.83E-04	+	1.56E-14	-	8.53E-05
Amidation	1.12E-01	+	2.28E-01	+	2.05E-01
Binding Site	4.81E-32	+	2.26E-110	-	7.86E-20
C-linked Glycosylation	1.68E-02	-	2.44E-01	-	1.09E-02
Crotonylation	2.89E-04	+	8.62E-07	-	2.69E-01
Gamma-carboxyglutamic acid	3.46E-08	+	1.80E-07	+	1.66E-02
Hydroxylation	4.42E-01	-	4.85E-01	-	4.90E-01
Methylation	2.24E-25	+	2.86E-28	+	4.84E-03
Myristylation	4.85E-01	+	3.47E-01	-	2.26E-01
N-linked Glycosylation	4.53E-273	+	2.09E-118	+	3.95E-163
O-linked Glycosylation	8.62E-24	-	9.79E-22	-	7.98E-06
Other Lipid Modification	2.21E-01	-	5.03E-01	-	3.03E-01
Other Modified Residue	6.75E-04	+	5.24E-04	+	1.67E-01
Palmitoylation	6.88E-02	-	1.67E-01	-	1.73E-01
Phosphorylation	1.28E-110	-	6.85E-66	-	7.78E-47
Prenylation	4.94E-03	-	1.22E-02	-	1.64E-01
S-Nitrosylation	7.20E-05	-	2.24E-06	-	2.91E-01
Sulfation	5.15E-01	+	3.33E-01	-	3.84E-01
Sumoylation	7.41E-08	-	2.25E-03	-	8.66914E-07
Ubiquitylation	1.18E-136	-	1.57E-71	-	1.10915E-61

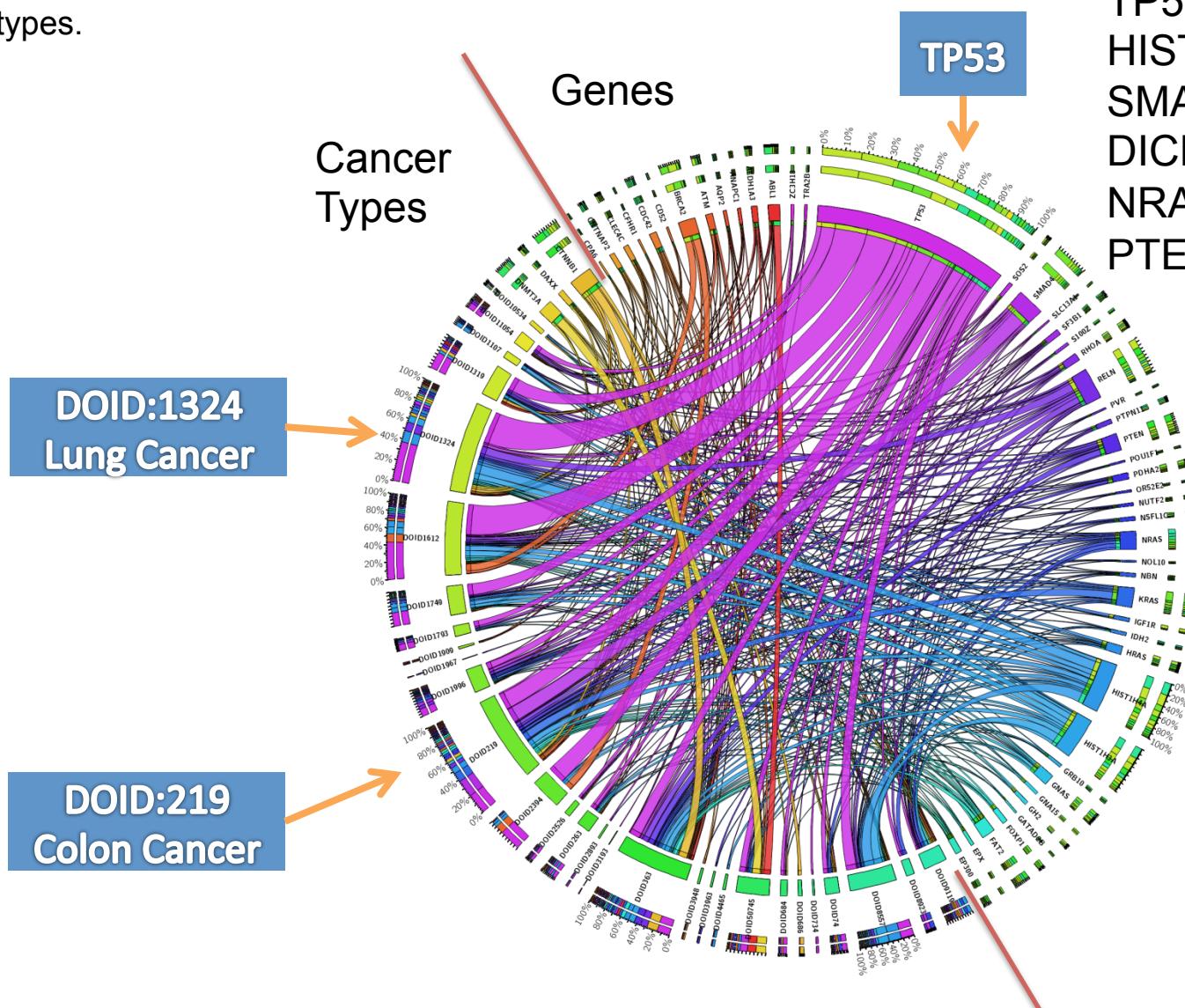
Pan-cancer analysis using HIVE



Creating functional profiles of cancers

990 cancer-associated mutations from 51 genes containing mutations that are across 3 or more cancer types.

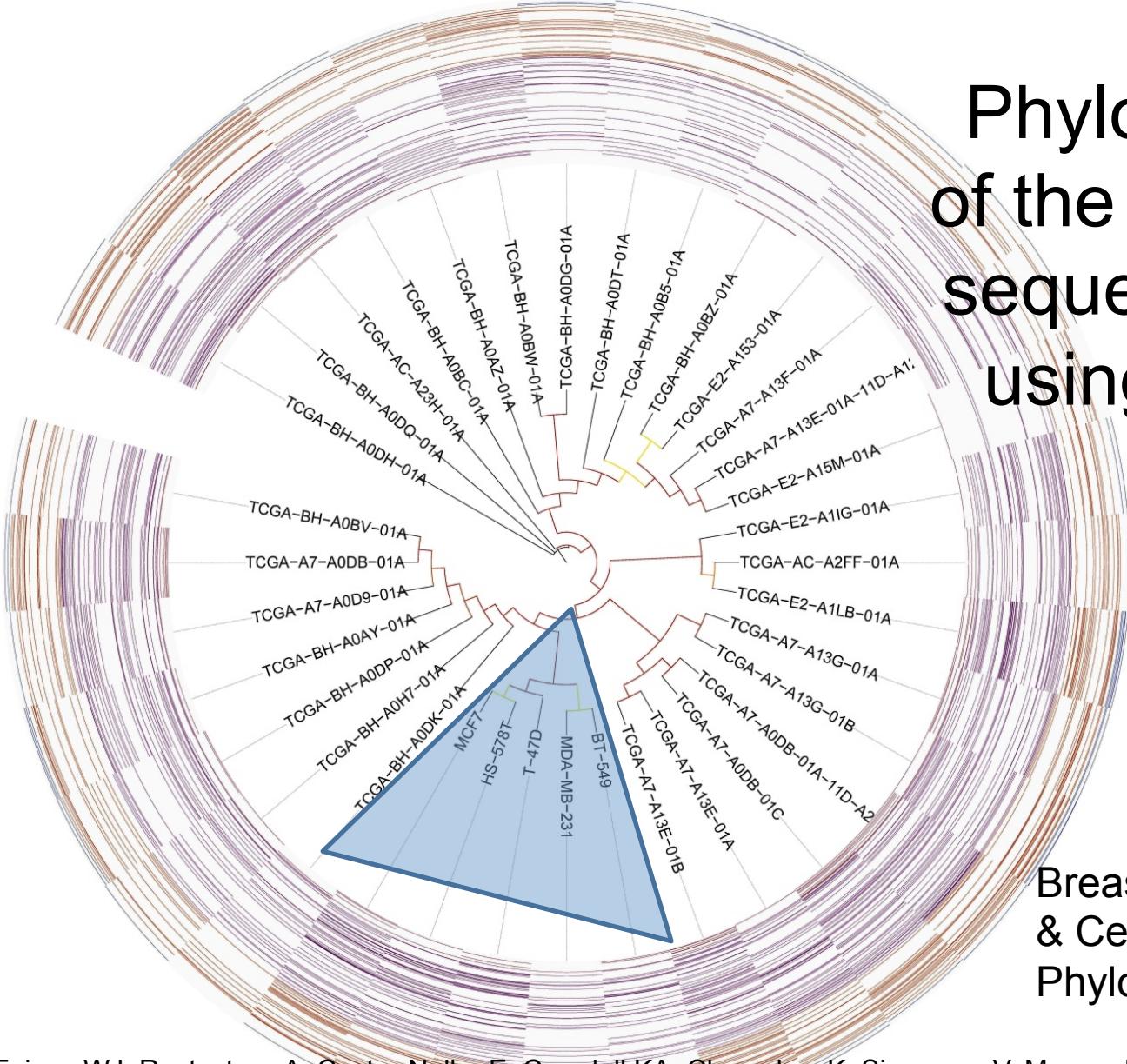
Results



Top 10 out of 51 key genes:
TP53, HIST1H4A, HIST1H3A, RELN, SMAD4, CTNN81, DICER1, KRAS, NRAS, BRCA2 and PTEN

Priority biomarkers
13 genes
106 mutations

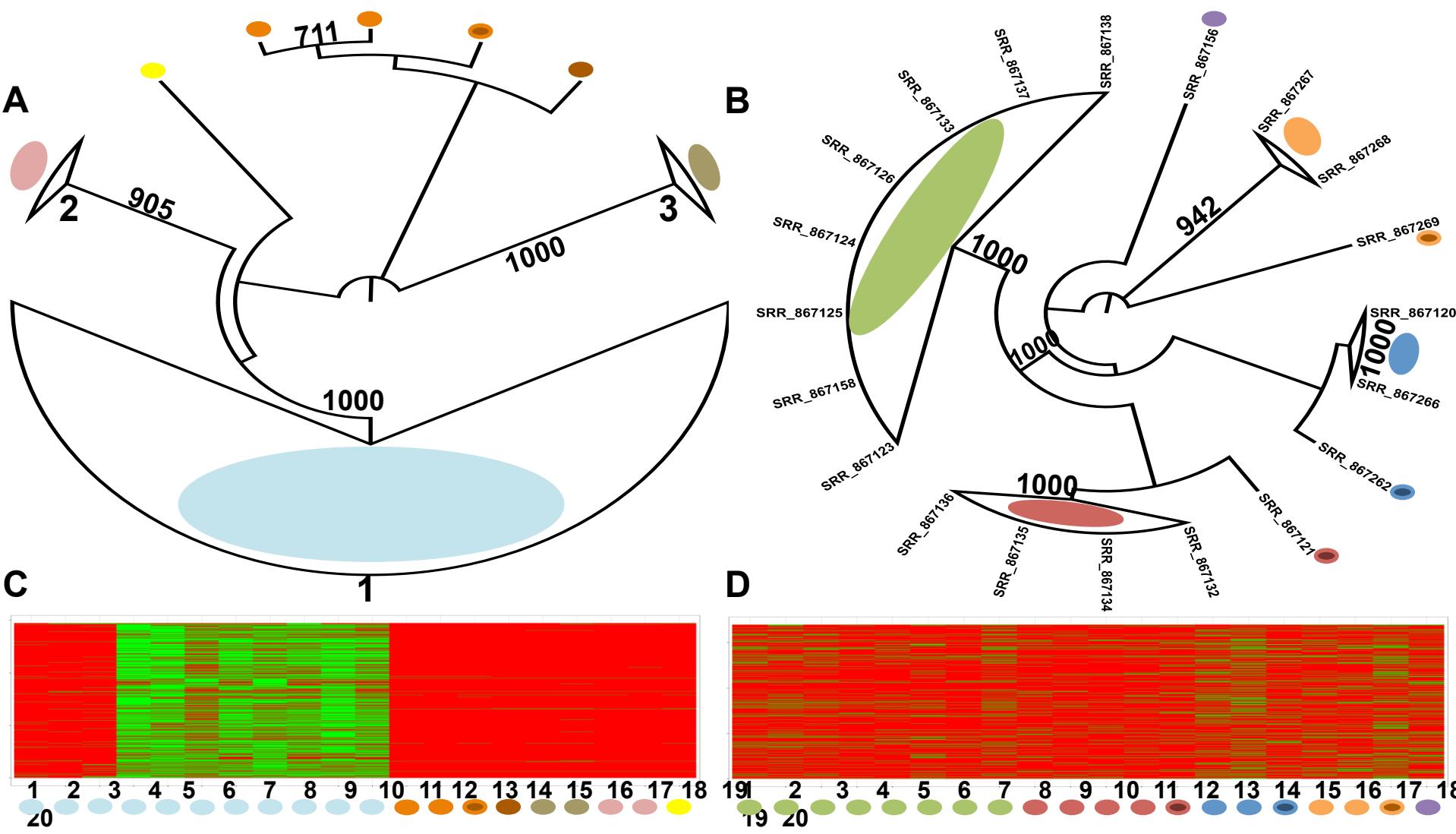
Phylogenetic tree of the whole exome sequencing results using PhyloSNP



Breast cancer tumor
& Cell-line sample exome-based
Phylogenetic analysis

Faison WJ, Rostovtsev A, Castro-Nallar E, Crandall KA, Chumakov K, Simonyan V, Mazumder R. Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. *Genomics*. 2014 Jul; 104(1):1-7.

Cole C, Krampis K, Karagiannis K, Almeida JS, Faison WJ, Motwani M, Wan Q, Golikov A, Pan Y, Simonyan V, Mazumder R. Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. *BMC Bioinformatics*. 2014 Jan 27;15:28.



We have the capability of performing whole genome
SNP-based phylogenetic analysis (phyloSNP)

CSR Breast Cancer Samples

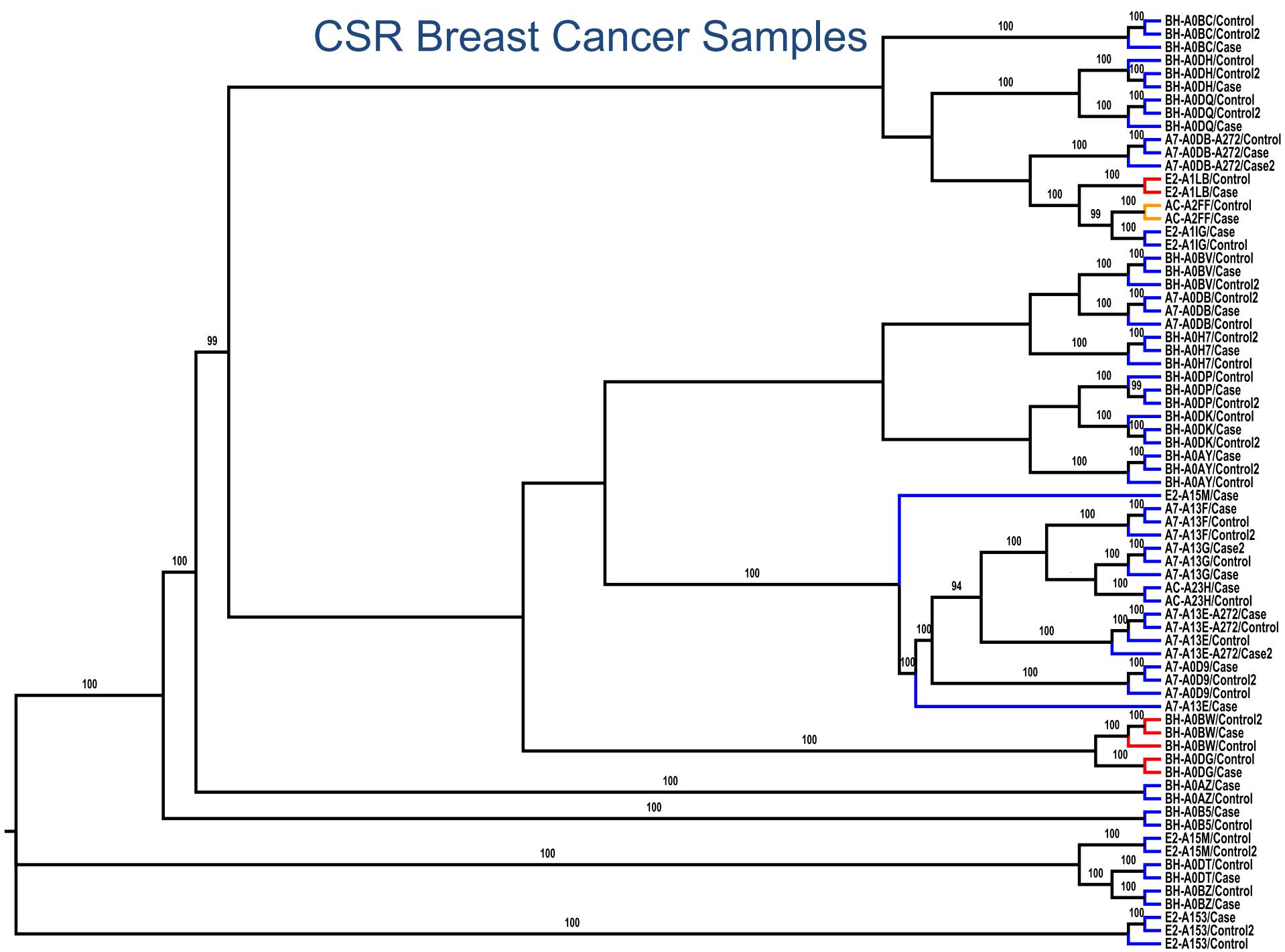
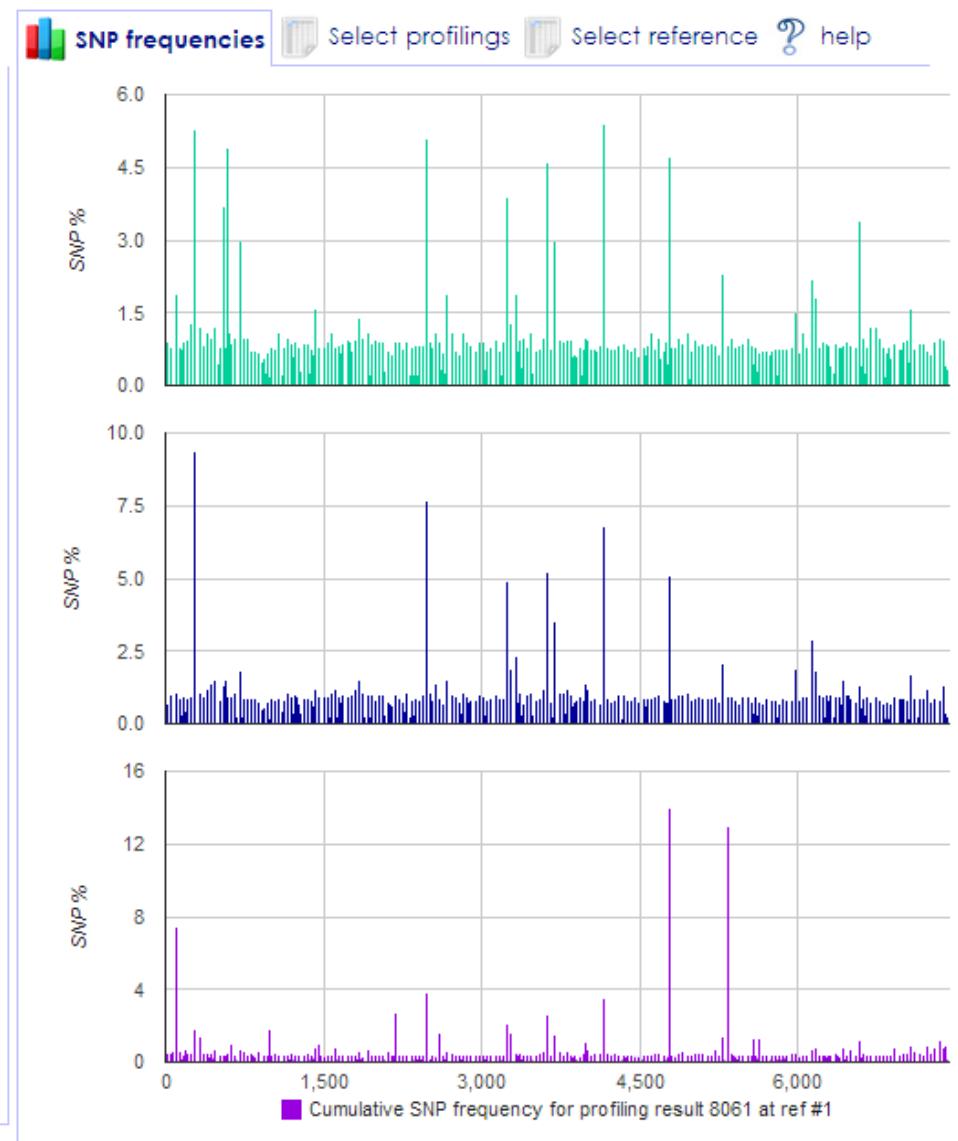
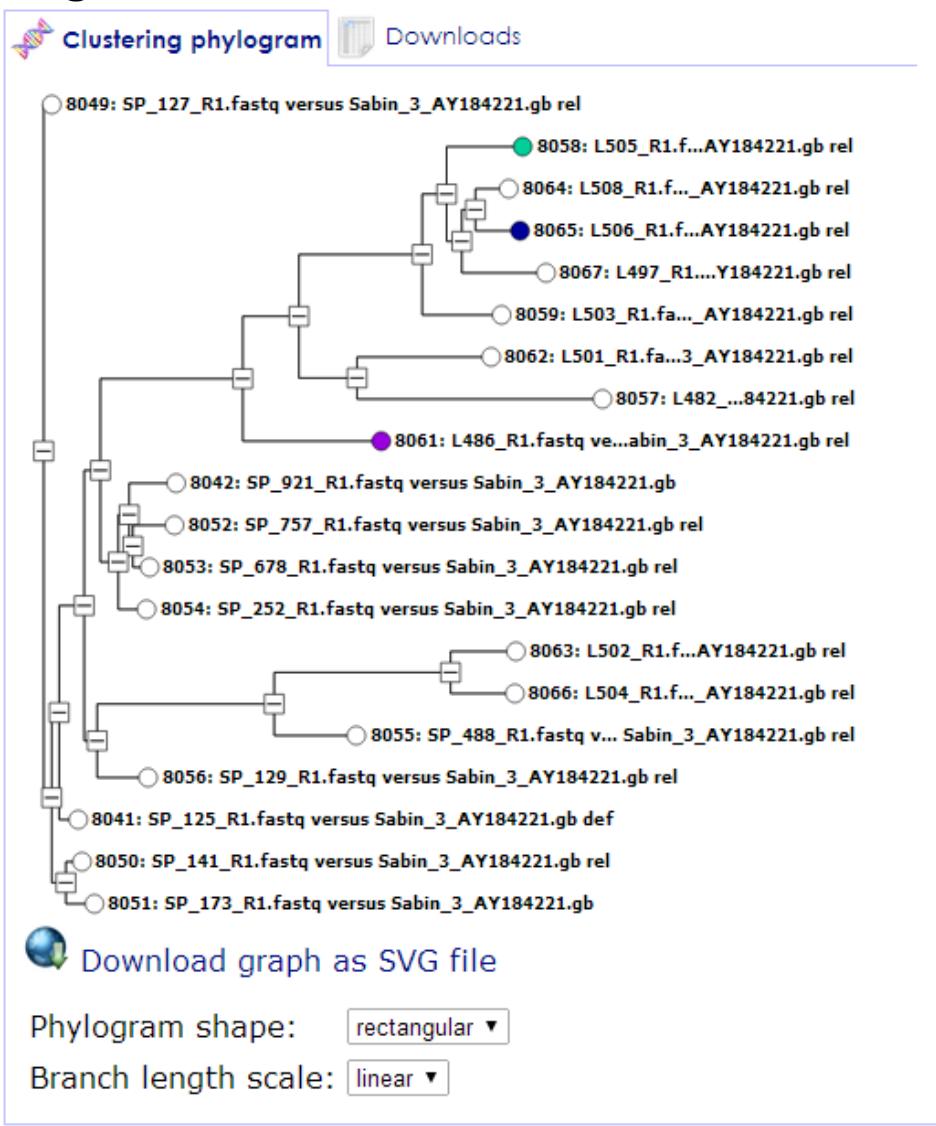


Figure 5



HIVE hierachal clustering

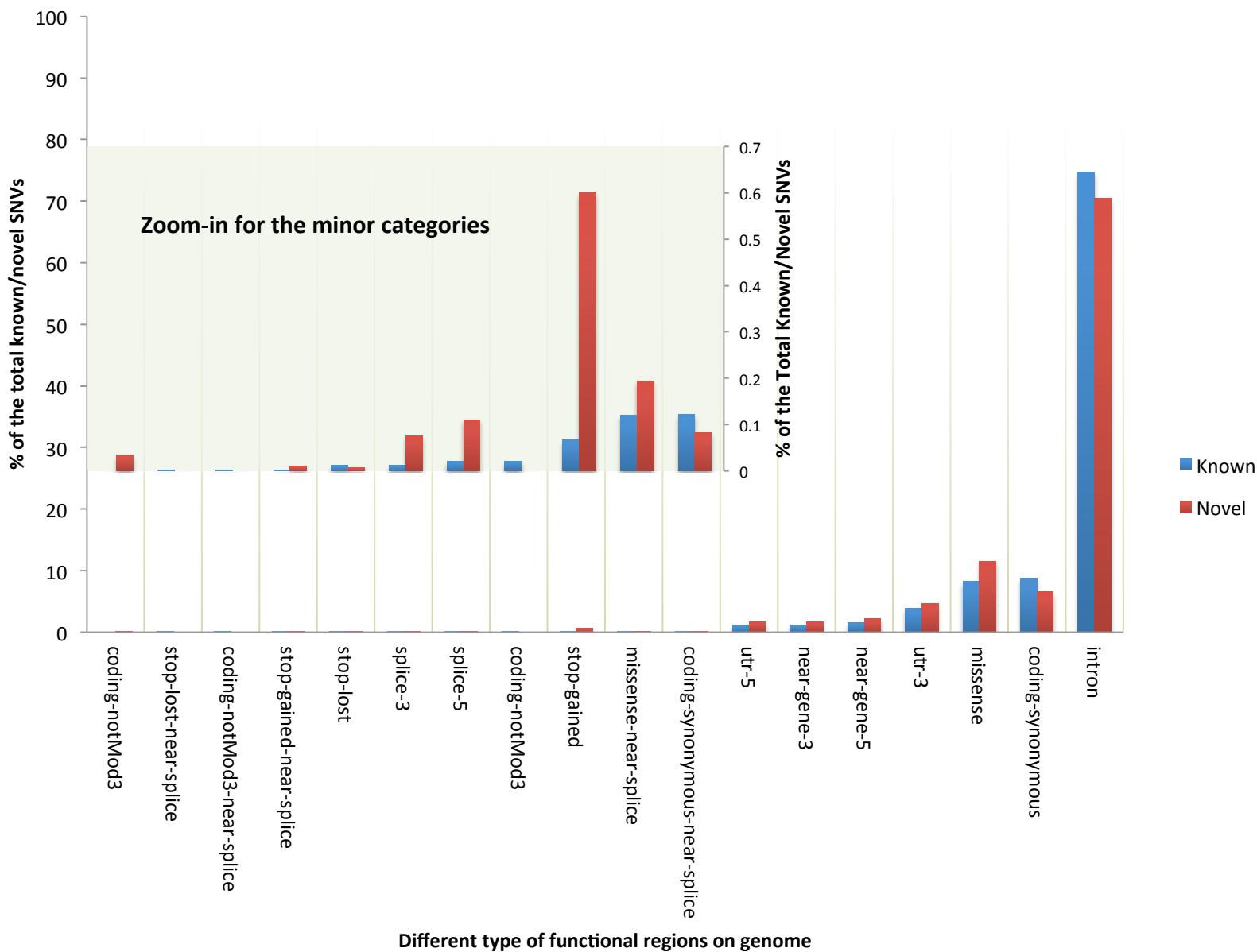
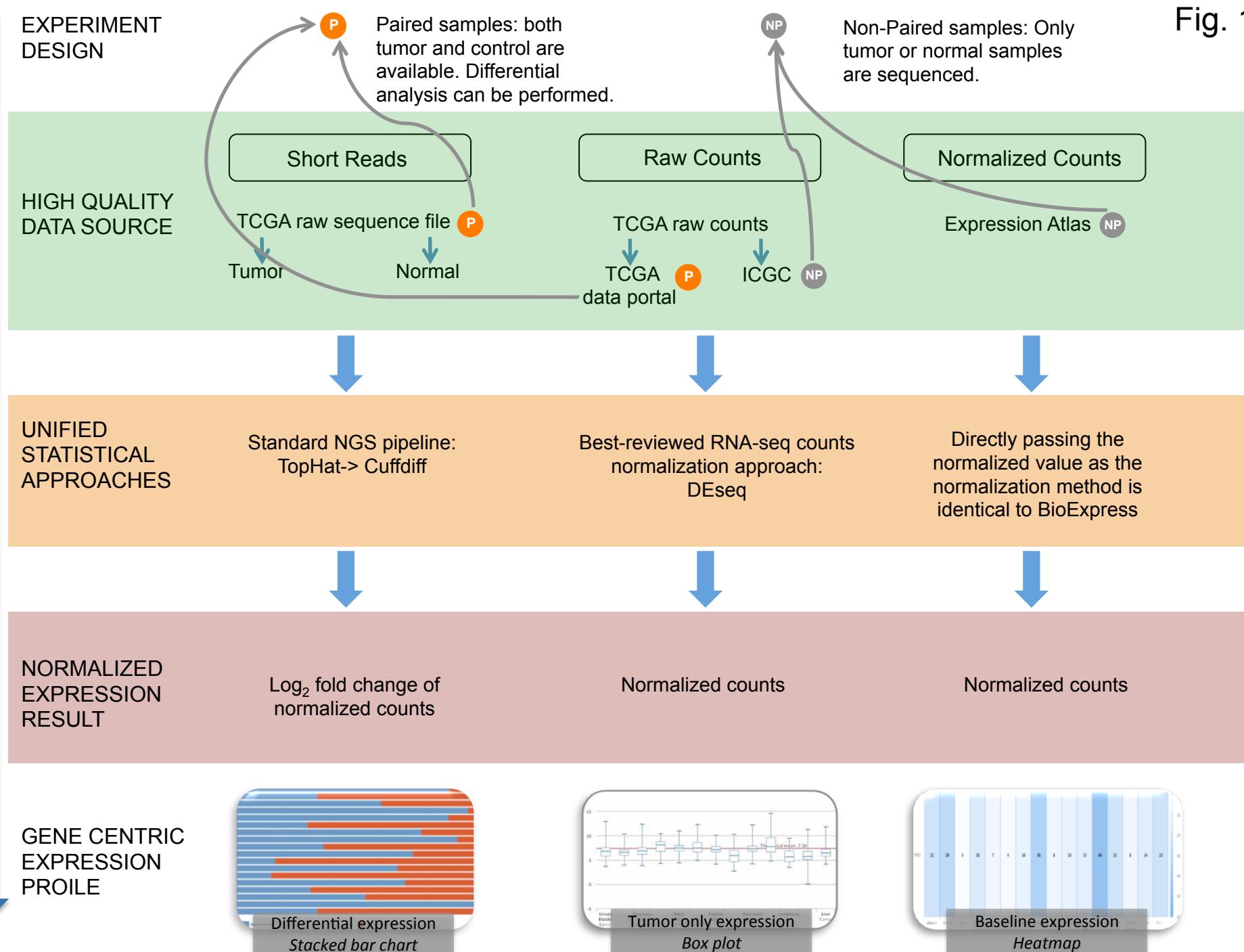


Fig. 1

EXPERIMENT DESIGN

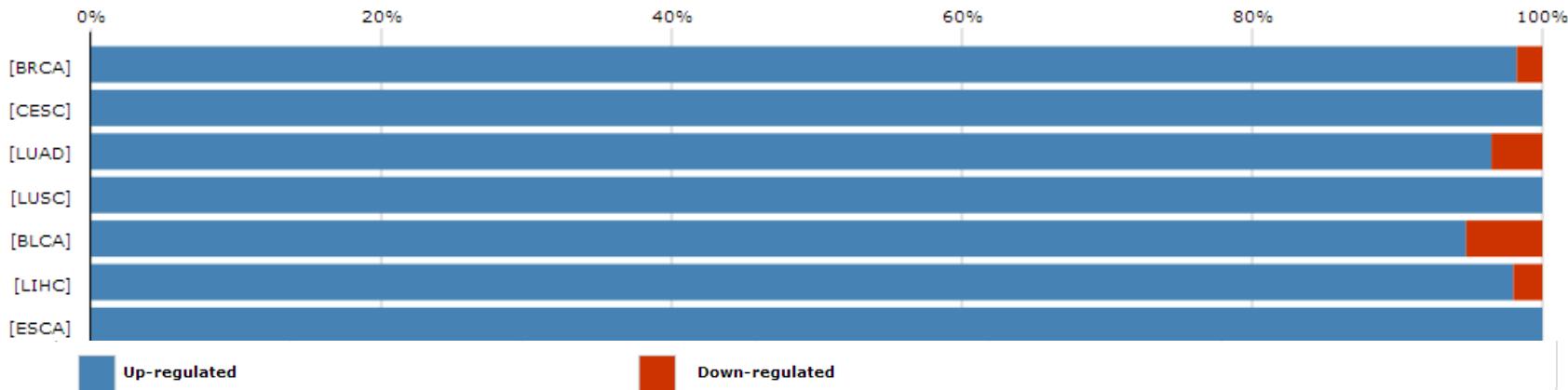


BioExpress v1.0 (beta)

TCGA/ICGC RNA-Seq based expression profiles

[HOME](#)[DIFFERENTIAL EXPRESSION](#)[TUMOR EXPRESSION](#)[BASELINE EXPRESSION](#)[ABOUT DATA](#)[Regulation/Freq.](#)[Significant/Freq.](#)

ASPM Expression Profile

[Table Column description](#) [Download](#)

972 Results for ASPM in BioExpress ▾

UniProtKB_AC	RefSeq	Gene	log2FoldChange	p_value	adjusted p_value	Significant	Regulated	Cancer Type	#Patients	Freq(sig%)	Sample ID	Data Source	PMID	Freq(up%)	Freq(Down%)
Q8IZT6	NP_001...	ASPM	6.79	8.23E-04	8.82E-02	Yes	Up	Lung adeno...	57	31.58	TCGA-38-4625	RNASeqV2	-	96.49	3.51
Q8IZT6	NP_001...	ASPM	9.45	2.39E-06	2.28E-03	Yes	Up	Breast invasi...	113	33.63	TCGA-BH-A18V	RNASeqV1	-	98.23	1.77
Q8IZT6	NP_001...	ASPM	5.19	4.92E-04	6.32E-02	Yes	Up	Breast invasi...	113	33.63	TCGA-E2-A1LH	RNASeqV1	-	98.23	1.77

Expression profile of a specific gene.

E.g. In figure below it shows that MUC16 is significantly regulated in Pancreatic adenocarcinoma

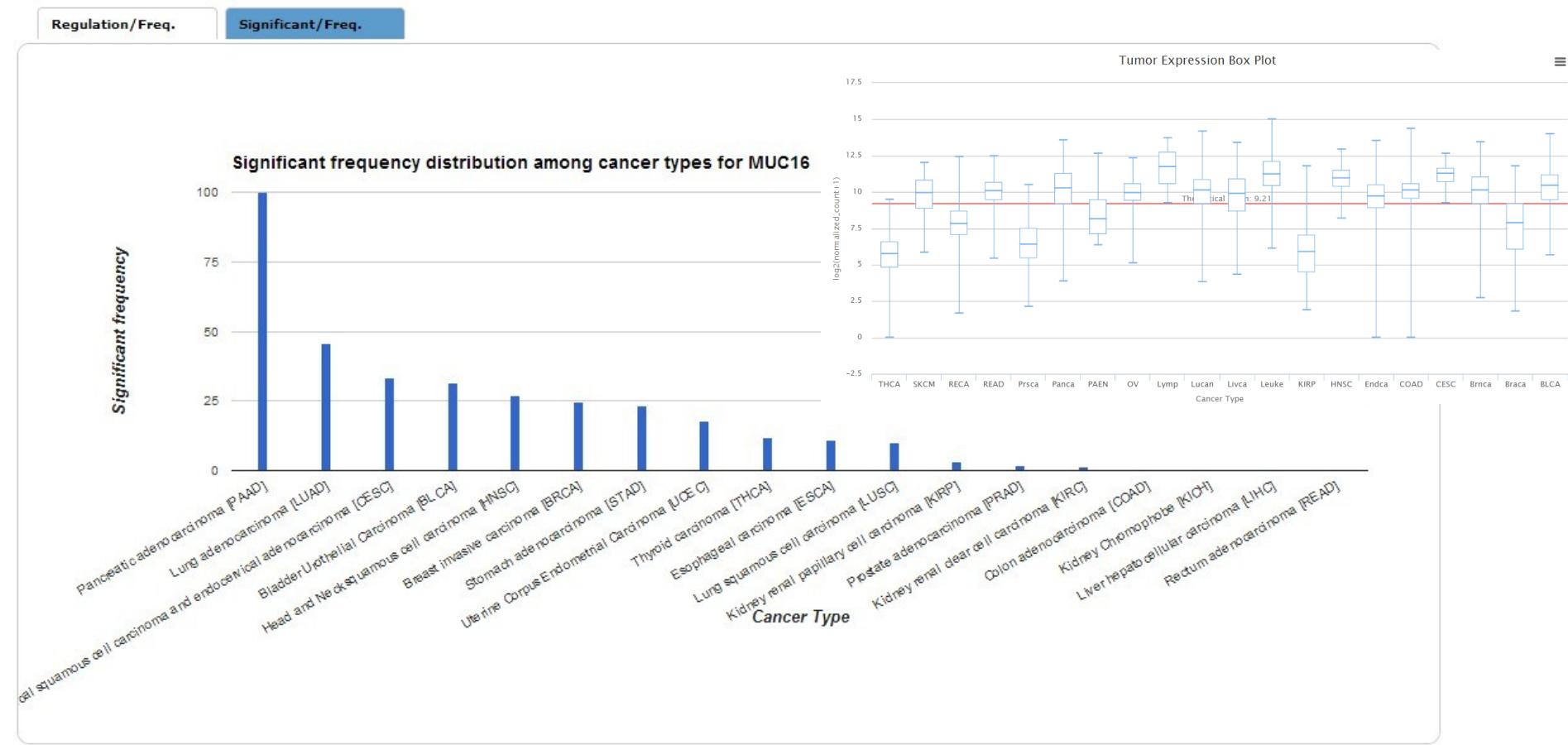
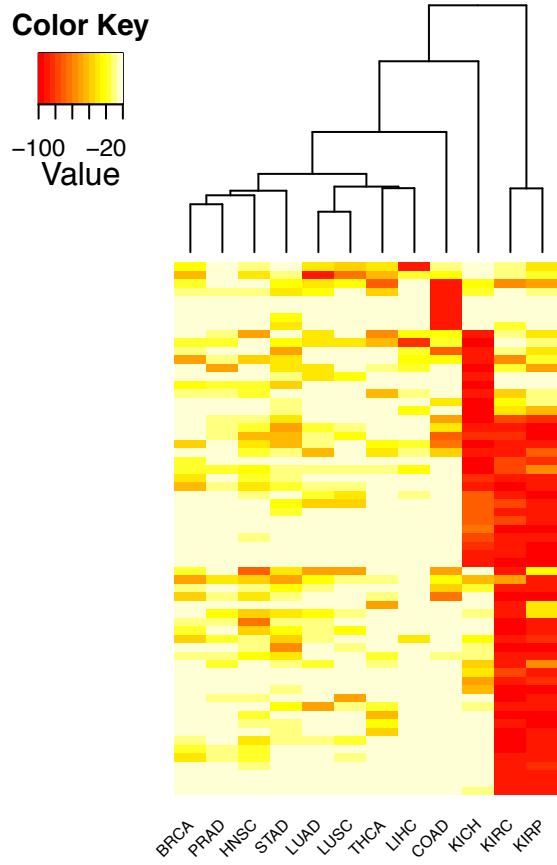
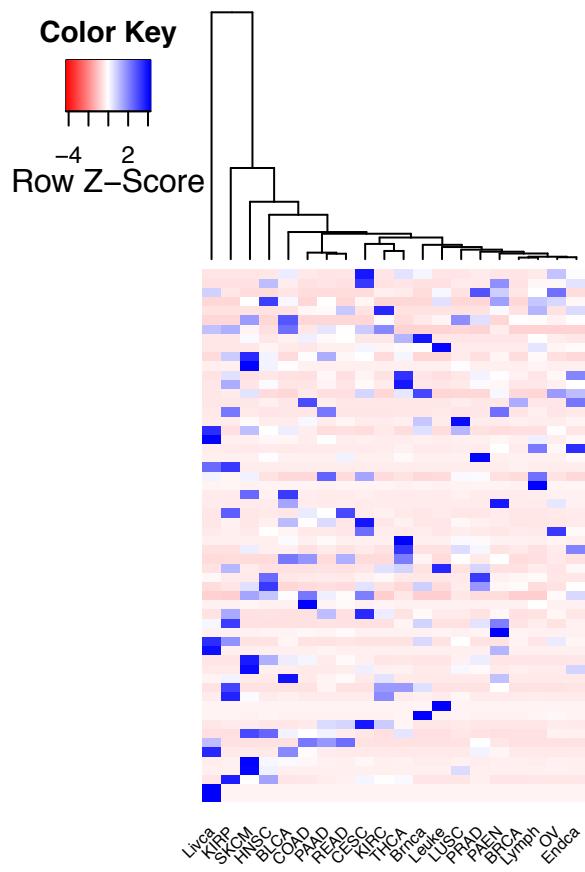


Fig. 3

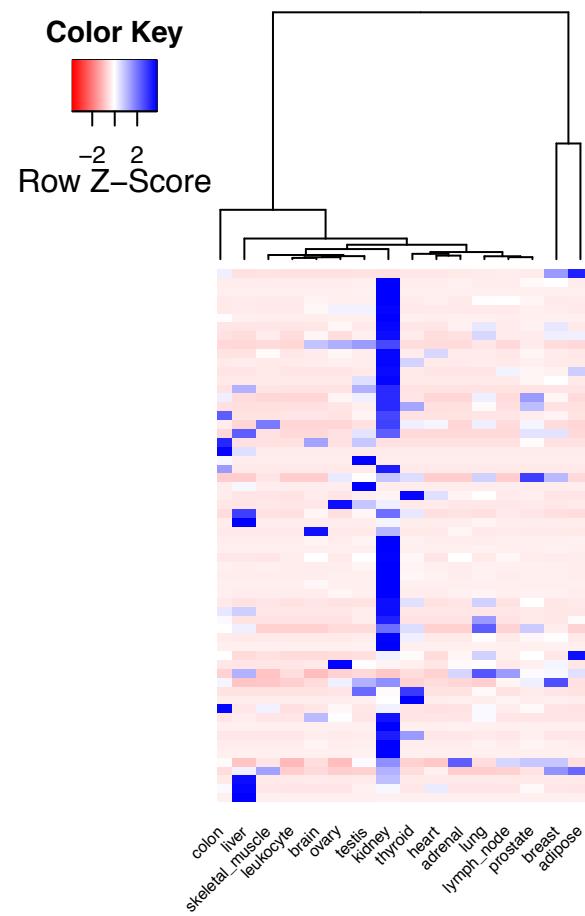
Panel A



Panel B



Panel C



Expression profile table

Click to sort

97 Results for MUC16 in BioExpress ▾



UniProtKB_AC	RefSeq	Gene	log2FoldChange	p_value	adjusted p_value	Significant	Regulated	Cancer Type	#Patients	Freq(sig%)	Sample ID	Data Source	PMID	Freq(up%)	Freq(Down%)
Q8WXI7	NP_078...	MUC16	6.18	1.16E-04	7.74E-03	Yes	Up	Pancreatic ...	3	100.00	TCGA-H6-A45N	RNASeqV2	-	100.0	0.0
Q8WXI7	NP_078...	MUC16	4.7	1.68E-06	3.24E-04	Yes	Up	Pancreatic ...	3	100.00	TCGA-H6-8124	RNASeqV2	-	100.0	0.0
Q8WXI7	NP_078...	MUC16	9.43	2.11E-10	1.28E-07	Yes	Up	Pancreatic ...	3	100.00	TCGA-HV-A5A3	RNASeqV2	-	100.0	0.0
Q8WXI7	NP_078...	MUC16	4.0	4.02E-05	9.59E-03	Yes	Up	Lung aden...	57	45.61	TCGA-49-6744	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	-0.58	6.84E-01	1.00E+00	No	Down	Lung aden...	57	45.61	TCGA-49-4512	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	5.95	7.20E-06	4.25E-03	Yes	Up	Lung aden...	57	45.61	TCGA-49-6745	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	6.45	2.28E-07	2.74E-04	Yes	Up	Lung aden...	57	45.61	TCGA-44-6147	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	2.87	2.11E-02	5.02E-01	No	Up	Lung aden...	57	45.61	TCGA-55-6982	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	0.1	9.57E-01	1.00E+00	No	Up	Lung aden...	57	45.61	TCGA-44-6778	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	3.4	6.47E-02	8.01E-01	No	Up	Lung aden...	57	45.61	TCGA-55-6971	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	-1.19	6.11E-01	1.00E+00	No	Down	Lung aden...	57	45.61	TCGA-55-6969	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	8.06	2.17E-03	1.38E-01	No	Up	Lung aden...	57	45.61	TCGA-50-5931	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	6.12	8.50E-07	2.86E-04	Yes	Up	Lung aden...	57	45.61	TCGA-49-6742	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	3.59	4.81E-05	4.69E-03	Yes	Up	Lung aden...	57	45.61	TCGA-44-6777	RNASeqV1	-	75.44	24.56
Q8WXI7	NP_078...	MUC16	6.27	2.28E-07	2.20E-04	Yes	Up	Lung aden...	57	45.61	TCGA-55-6970	RNASeqV1	-	75.44	24.56

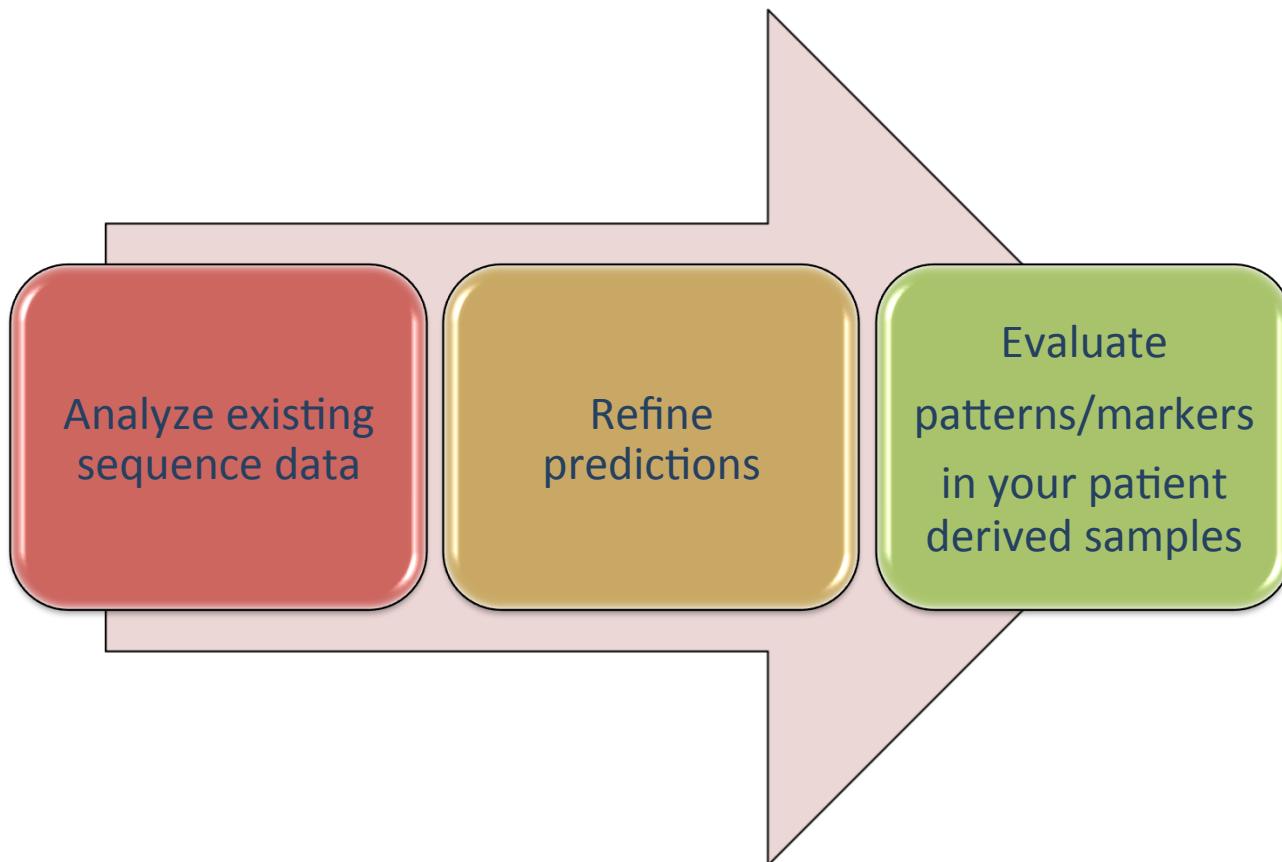
Download table in excel format

Genes differentially expressed in all cancers

Gene	UniProtKB AC	Protein Name	Localization	PTM
CCL21	O00585	C-C motif chemokine 21	Secreted	Glyco regulation
GGT6	Q6P531	Gamma-glutamyltransferase 6	Transmembrane	Glycoprotein
UBD	O15205	Ubiquitin D	Cytoplasm/nucleus	Acetylation
MMP7	P09237	Matrilysin	Secreted	Glycosylation?
NCAM1	P13591	Neural cell adhesion molecule 1	Secreted	Phosphoprotein
CHRDL1	Q9BU40	Chordin-like protein 1	Secreted	Glycoprotein
WFDC2	Q14508	WAP four-disulfide core domain protein 2	Secreted	Glycoprotein
LCN2	P80188	Neutrophil gelatinase-associated lipocalin	Secreted	Glycoprotein
KRT80	Q6KB66	Keratin, type II cytoskeletal 80	Cytoplasm	Phosphoprotein

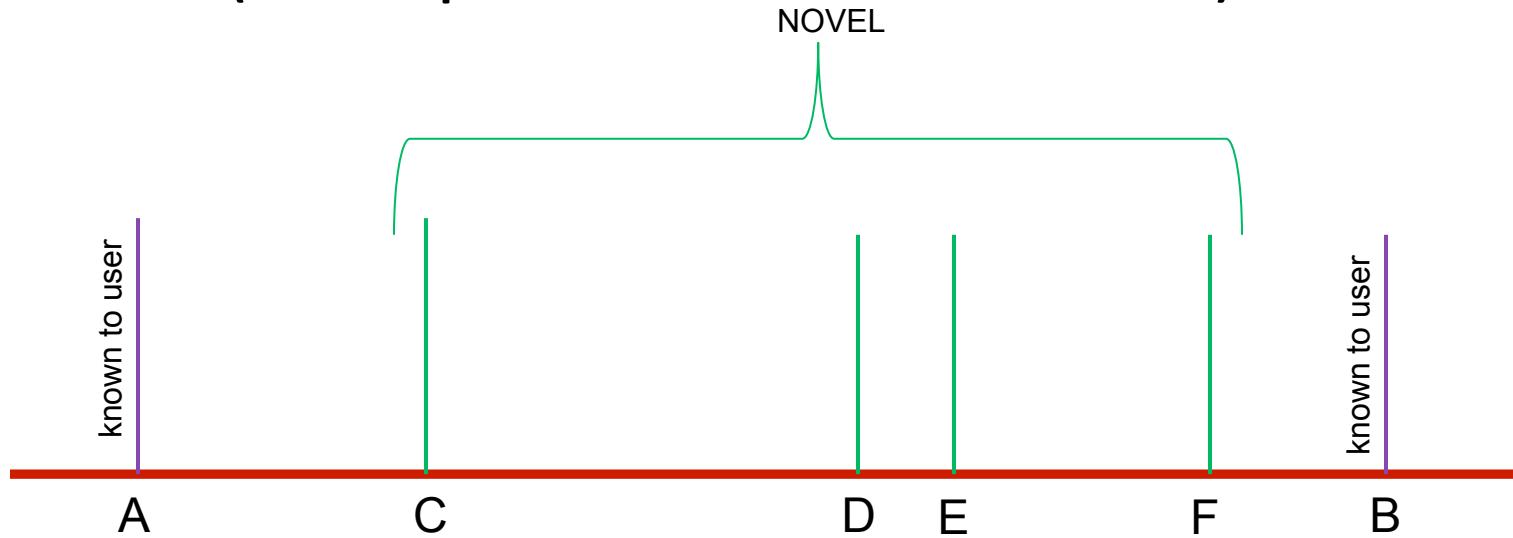
Why TCGA/ICGC etc.?

- **STOP** before you sequence



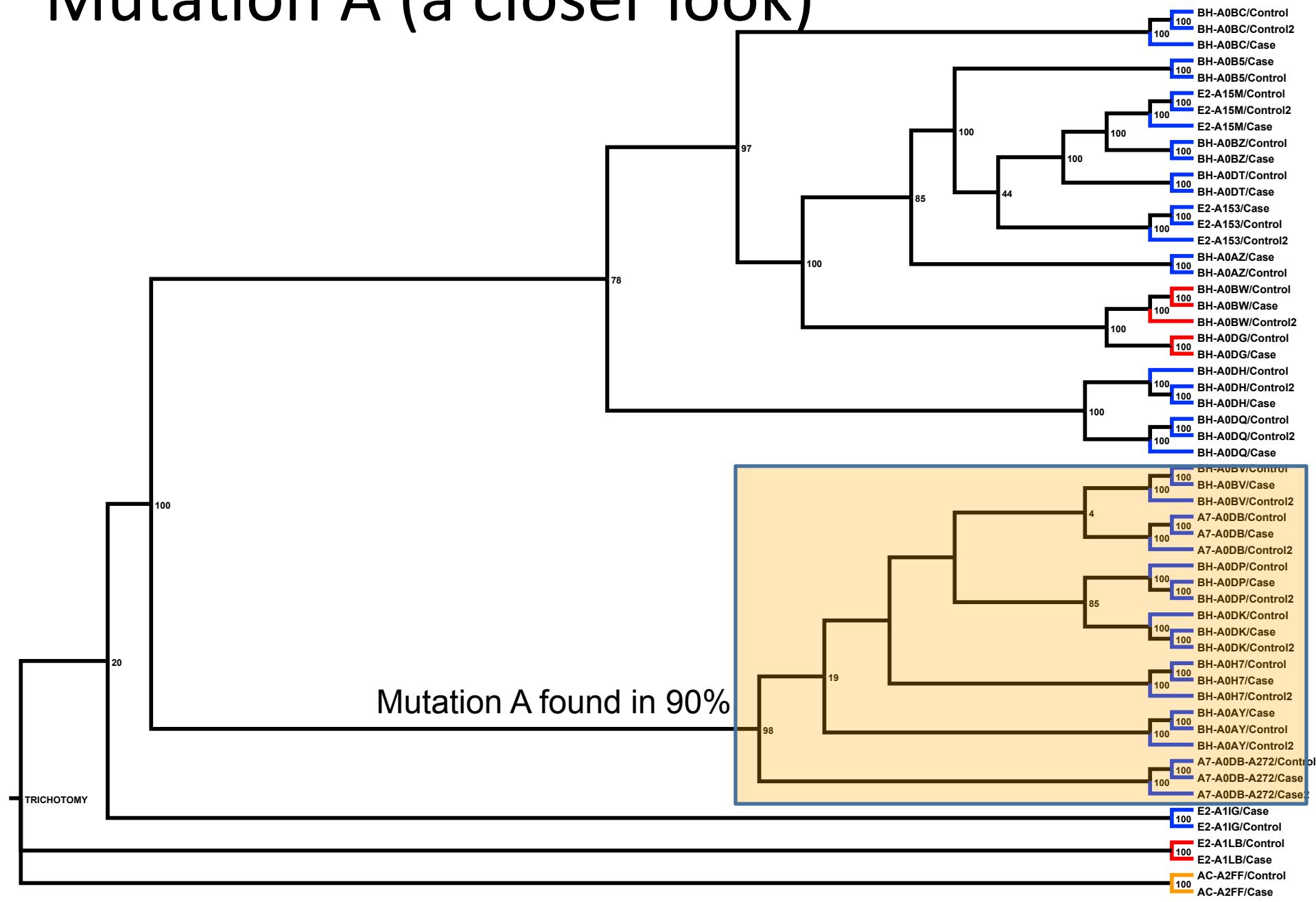
Mutation profiling

(Example similar to real data)



M	Position	Category	SNP ID	Effect	Found in
A	A	2	rs10235	DISUL	TCGA/BRCA
B	B	30	COS123	PTM	COSMIC
C	C	5	-	CONS	TCGA/BRCA
D	D	10	-	CONS	TCGA/BRCA
E	E	90	-	PHOS	TCGA/BRCA
F	F	10	rs13456	-	TCGA/BRCA

Mutation A (a closer look)



Our recent pan-cancer analysis results

- **62 mutations in 13 genes** affecting functional sites such as DNA, ATP binding and various PTM sites occur across several cancers and can be prioritized for additional evaluations and investigations. Nucleic Acids Research. Accepted. *Human germline and pan-cancer variomes and their distinct functional profiles.*

Gene	Variation	Functional Site	Cancer Type	Conserv.	PDB ID
TP53	K164E	Acetylation	ovarian cancer; brain cancer; colon cancer; lung cancer	Yes	3D06
SF3B1	K700E	Ubiquitylation	breast cancer; diffuse large B-cell lymphoma; acute myeloid leukemia; pancreatic cancer; prostate adenocarcinoma	Yes	NA
...

HIVE Genecast app

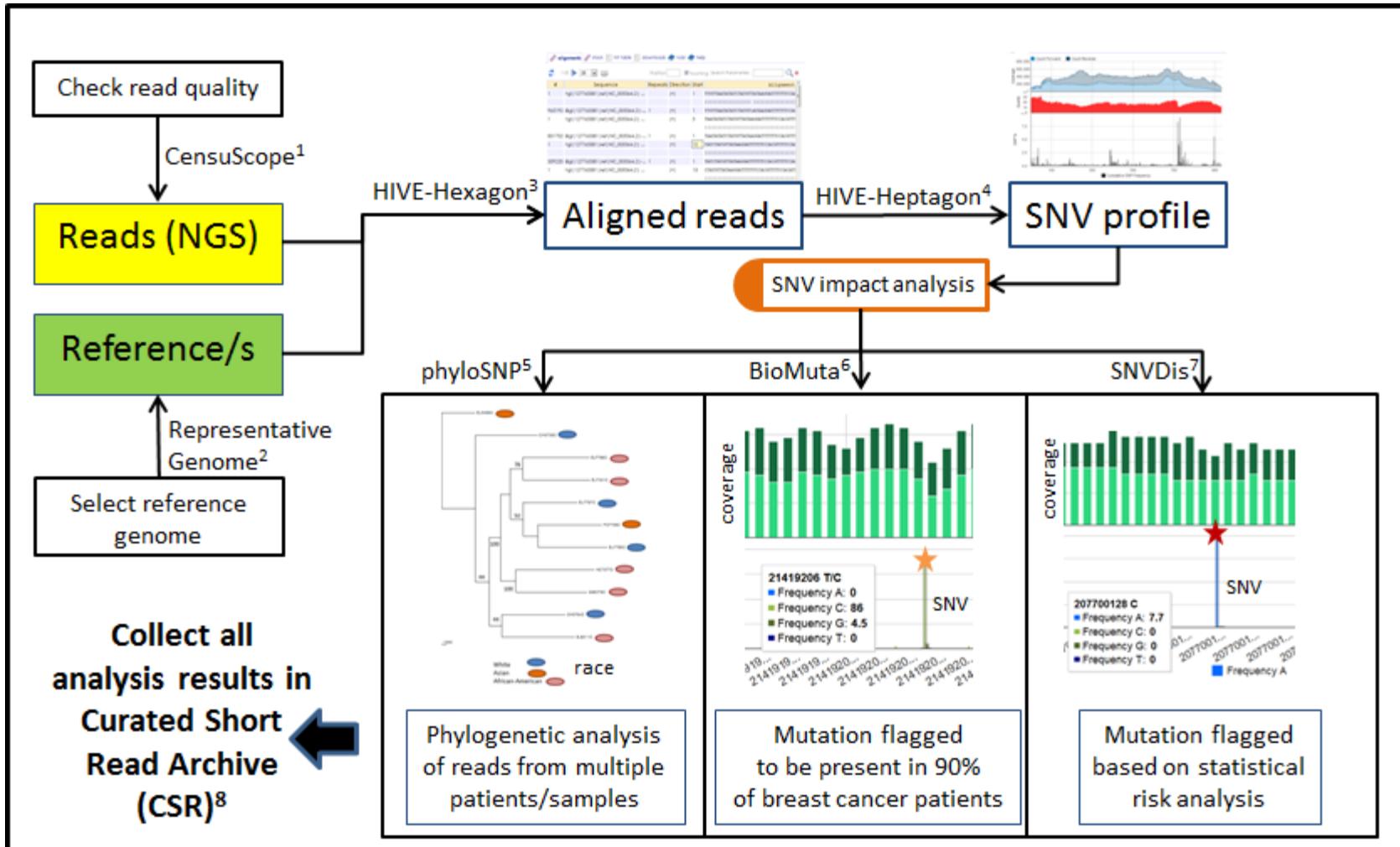
Tools for researchers and biocurators

<http://hive.biochemistry.gwu.edu/dna.cgi?cmd=hivegenecast>

The figure displays five screenshots of the HIVE Genecast mobile application interface, showing its various features:

- BioMuta Screen:** Shows a search bar for HGNC gene symbols. A search for "MUC16" has been entered, and a "Submit" button is visible.
- BioExpress Screen:** Shows a search bar for HGNC gene symbols. A search for "MUC16" has been entered, and a "Submit" button is visible.
- Detailed Mutation Data Screen:** Displays a table of mutations for the gene MUC16. The table includes columns for Pos(A), Ref(A), Var(A), Damaging, PMID, Cancer, and Source. One row is highlighted with a yellow background.
- HIVE Genecast Screen:** Shows a detailed view of a mutation for MUC16. It includes fields for Gene (MUC16), UniProtKB (Q8WXI7), RefSeq (NM_024690.2), SNV Position (chr19:9090928-9090928), Pos(N) (1091), Ref(N) (G), Var(N) (A), Pos(A) (296), Ref(A) (S), Var(A) (L), Polyphen Pred. (possibly damaging), PMID (22980975), Cancer Type (Lung adenocarcinoma [LUAD]), and Source (COSMIC).
- HIVE Pubcast Screen:** Shows a news article titled "A transcriptional switch underlies commitment to sexual development in malaria parasites." It includes details about the date (2/28/2014), journal (Nature), and a brief summary of the research findings.

Workflows



Future plans

- Focused analysis/curation GlycoMuta/
GlycoExpress pilot database currently hosts data
from 32 genes requested by Alliance members.
- Closer integration with EDRN portal (CDE/
Ontology)
- Update existing data and add more data as they
become available
- Provide mechanism for EDRN members to scan
TCGA and ICGC genomic data using HIVE
- Data interpretation, biocuration and literature
mining

Acknowledgements

**HIVE TEAM MEMBERS, COLLABORATORS, USERS
&
BIOCURATORS (CDD, UniProt, RefSeq, COSMIC and many more)**

Special thanks: Vahan Simonayn (FDA) Radoslav Goldman (GU), Dan Crichton (JPL), Karl Krueger (NCI) & Sudhir Srivastava (NCI)

Contact

mazumder@gwu.edu

Funding

Alliance of Glycobiologists (subaward)
EDRN (Associate Membership)
GWU start-up & internal funding
FDA/ORISE fellowships

<http://hive.biochemistry.gwu.edu/tools/biomuta2beta/>

<http://hive.biochemistry.gwu.edu/tools/bioexpress/>

<http://hive.biochemistry.gwu.edu/dna.cgi?cmd=phylosnp>

<http://hive.biochemistry.gwu.edu/help/HIVEWwhite29Jan14.pdf>